Submitted Solutions to the Banff Challenge 2a Problems

T. Junk[1], W. Fisher[2], J. Linnemann[2], R. Lockhart[3], L. Lyons[4]
[1]*Fermi National Accelerator Laboratory*
[2]*Michigan State University*,
[3]*Simon Fraser University*
[4]*Imperial College London*

February 15, 2011

Abstract

The workshop at the Banff International Research Station, 10w5068, on "Statistical Issues Relevant to Significance of Discovery Claims", raised several interesting issues that are best illustrated with concrete examples that participants can try out and discuss the issues that arise. Many solutions were offered by the participants exploring several different techniques to solve the problems, an all solutions are of very high quality. The challenge datasets are designed to test these methods for data for which the true values of the parameters remain unknown to the participants. This document provides a brief summary of each method provided by the participants, and details the performance of the solutions.

1 Introduction

The two problems are specified in a separate note, available at

`http://www-cdf.fnal.gov/~trj`

This document describes the methods used to generate the simulated datasets, and summarizes the entries received from participants, showing the performance of each. Each participant was required to provide a description of the methods used to solve the two problems, and these documents are available separately. All of the responses to Problem 1 are grouped together, are are those for Problem 2

2 Challenge Problem #1

2.1 Challenge Datasets

The challenge datasets were generated randomly according to the distribution described in the challenge note:

$$B(x) + S(x) = Ae^{-Cx} + De^{-(x-E)^2/2\sigma^2}. \quad (1)$$

There are 24 different subsets of the simulated pseudoexperiments, corresponding to different choices of D and E, and these are listed in Table 1. The numerical choices were governed by the desire to have a correct-discovery rate that can be measured accurately with a limited number of repetitions, and thus should not be too close to 0% or 100%, and that we would like to test more than one regime. Signals with large values of D and small values of E can presumably be treated with a Gaussian approximation, while signals with small values of D and large values of E are very sensitive to the Poisson nature of the data in sparsely populated areas of the distribution.

The parameters D and E are parameters of interest and are not affected by unknown values of nuisance parameters. This is somewhat unrealistic, since in a real high-energy physics context physicists are usually not entirely sure of the fraction of collisions that would trigger our detectors' readouts and pass our event selection requirements, although we have estimates of these numbers. Similarly, the location of a peak does not always correspond to the true value of the mass of a new particle, although the significance of a peak should not be affected by the uncertainty in the relationship between the measured peak position and the underlying process that makes events in the peak. Similarly, the trigger and event selection acceptance uncertainty should have little impact on the significance of a peak that is found, although they do have impacts on the expected sensitivity, signal rate measurements, and limits.

The background parameter A was chosen for each simulated dataset from its prior distribution, a Gaussian centered on 10000 with a width of 1000. An integer n_b was then drawn from a Poisson distribution whose mean is the total background integral from $x = 0$ to 1 using the randomly selected value of A. Then n_b marks x were generated from the exponential distribution $B(x)$. A similar procedure was followed for generating marks for the signal component, according to $S(x)$. The marks were then shuffled and written out to the challenge dataset file. Simulated datasets from the 24 categories were also shuffled so that no clue to the injected values would be provided by the ordering of either the datasets or the marks within a dataset.

The presence of a nuisance parameter in the null and test hypotheses complicates the definition of the Type-I error rate. One approach is to evaluate the Type-I error rate as a function of the true value of the unknown nuisance parameter(s). Another approach is to evaluate the Type-I error rate in the prior-predictive ensemble whose generation is described above. A third is to quote the largest Type-I error rate for a fixed range of values of the nuisance parameters. The ideal that a method should cover for all values of the nuisance parameter

requires a specification of what is meant by "all". The approach here is to quote the error rate and the correct-discovery rates using the prior-predictive ensemble, although this is not the only valid definition. A method which has a Type-I error rate which is larger than the stated value, which is usually written in a high-energy physics publication as a confidence level or a significance level, is said to undercover and is unlikely to pass collaboration review.

A feature of challenge Problems 1 and 2 is that signal rate intervals were requested only in the case that evidence is claimed, and the problem statement asks for zero to be entered if evidence is not claimed. These instructions reflect a flip-flopping procedure which is very commonly used in particle physics. If a particle physics collaboration measures the mass of a new particle but does not claim evidence for the new particle, the result may be easily misconstrued.

Nonetheless, not quoting the measured signal yield in simulated datasets for which evidence is not claimed biases upwards the measured signal yields and the intervals containing them. A simple example is the null hypothesis – the true signal rate is zero in null hypothesis simulated datasets, but in 1% of them, a method that is performing well should claim evidence for a signal. Even if the set of intervals for the signal rate cover properly for a method, selecting this sample of them will in general not have proper coverage. This is true to a lesser extent for test hypotheses with true signals present.

A final feature of Problem 1 is that at most one signal is present, at a single value of E. In a real experiment in which the signal is *a priori* unknown, there may be more than one signal present. Since most methods fit for the background rate in the process of testing for the signal, a second signal (or more) will change the background fit. One may legitimately ask whether all of the events are signal events from a broad spectrum of multiple signals, and this is where some theoretical input and auxiliary information from other experiments is needed to constrain the background. For this problem, we treat the presence of at most one signal as auxiliary *a priori* information. The challenge datasets were generated with no more than one signal in each.

2.2 Solutions Received

Table 2 lists the contributors who provides solutions to Problem 1, the fractions of null-hypothesis simulated datasets that resulted in a discovery claim, and the fractions of simulated datasets that were in the power test samples that resulted in discovery claims, compared with the estimations provided by the participants. In high-energy physics experiments, the claimed power is quite important – it plays a pivotal role in deciding which experiments to fund, which to give extended running time to, and it plays a key role in individual collaborators' decisions of which topics to pursue within a running experiment. It is vital to be able to compute these numbers reliably, and Banff Challenge 2 is an ideal forum in which to test these computations. Methods should have a Type-I error rate not exceeding 1%, the specified level for this exercise.

Table 1: Problem 1 challenge dataset categories, listing the input values of E and D, the signal peak position and the signal rate parameters, respectively. The first category is the null hypothesis.

Category	E_{input}	D_{input}	n_{rep}
1	—	0.00	15400
2	0.50	83.78	200
3	0.38	265.96	200
4	0.10	1010.65	200
5	0.10	478.73	200
6	0.66	66.49	200
7	0.78	39.89	200
8	0.10	744.69	200
9	0.50	136.97	200
10	0.90	15.29	200
11	0.50	190.16	200
12	0.14	664.90	200
13	0.50	163.57	200
14	0.38	531.92	200
15	0.14	1196.83	200
16	0.50	110.37	200
17	0.10	1276.62	200
18	0.90	20.61	200
19	0.66	132.98	200
20	0.90	12.63	200
21	0.90	17.95	200
22	0.90	23.27	200
23	0.78	79.79	200
24	0.10	1542.58	200

Table 2: Listing of the estimated and measured correct-discovery rates for the three scenarios of Problem 1. The SCT's claimed discovery rate for the third scenario is probably a typo. Stefan Schmitt states that his unbinned sensitivities are rather similar to his binned sensitivities.

Contributor	Type-I Error Rate	$D = 1010, E = 0.1$		$D = 137, E = 0.5$		$D = 18, E = 0.9$	
	Measured	Claimed	Measured	Claimed	Measured	Claimed	Measured
Tom Junk	0.0097 ± 0.0008	0.256	0.3150 ± 0.0328	0.543	0.6100 ± 0.0345	0.108	0.1350 ± 0.0242
Wolfgang Rolke	0.0103 ± 0.0008	0.356	0.3800 ± 0.0343	0.457	0.5250 ± 0.0353	0.184	0.2150 ± 0.0290
Stanford Challenge Team (SCT)	0.0077 ± 0.0007	0.3483	0.3550 ± 0.0338	0.4335	0.5200 ± 0.0353	0.0175	0.2100 ± 0.0288
Eilam Gross & Ofer Vitells	0.0082 ± 0.0007	0.35	0.3600 ± 0.0339	0.46	0.5250 ± 0.0353	0.19	0.2100 ± 0.0288
Valentin Niess	0.0111 ± 0.0008	0.603	0.3250 ± 0.0331	0.87	0.5300 ± 0.0353	0.12	0.1950 ± 0.0280
Georgios Choudalakis	0.0110 ± 0.0008	0.213	0.1600 ± 0.0259	0.290	0.3500 ± 0.0337	0.107	0.1300 ± 0.0238
Mark Allen	0.0106 ± 0.0008	0.385	0.4000 ± 0.0346	0.486	0.5250 ± 0.0353	0.187	0.2100 ± 0.0288
Frederik Beaujean (BAT)	0.0000 ± 0.0000		0.0000 ± 0.0000		0.0300 ± 0.0121		0.0050 ± 0.0050
Stefan Schmitt							
Unbinned	0.0112 ± 0.0009		0.4500 ± 0.0352		0.5450 ± 0.0352		0.1850 ± 0.0275
Binned	0.0110 ± 0.0008	0.37	0.3850 ± 0.0344	0.53	0.5450 ± 0.0352	0.17	0.2200 ± 0.0293
Stefano Andreon							
$p < 3 \times 10^{-3}$	0.0126 ± 0.0013		0.4811 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120
$p < 4 \times 10^{-3}$	0.0191 ± 0.0016		0.5189 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120

2.2.1 From Tom Junk

For Challenge Problem #1, Tom Junk provided a solution based on an unbinned profile likelihood test statistic. Two fits are done, both using MINUIT, one in the test hypothesis, and one for the null hypothesis. Simulated datasets were generated using the prior-predictive ensemble. The Look-Elsewhere Effect is incorporated by testing all datasets in the same way, allowing a peak to be found anywhere in the ranges $0 < E < 1$ and $0 < D$. Tom reports the values of D and E returned by the MINUIT fit.

Table 3 lists the error rates in the challenge datasets for Tom's solution. The Type-I error rate is just under 1% as desired, although the confidence intervals for the fitted signal in those datasets with a Type-I error do not contain zero signal very often. One does not expect that, as they are 68% intervals and we insist that only 1% of outcomes have a Type-I error.

Table 3: Problem 1 performance evaluation for Tom Junk's solution. The columns are $E_{\rm true}$, the input value of the peak position, $D_{\rm true}$, the input value of the signal rate parameter, $n_{\rm rep}$, the number of simulated datasets in the 20,000 sample in this category, $n_{\rm disc}$, the number of datasets on which a discovery was reported, $f_{\rm disc}$, the fraction of datasets on which a discovery is reported, $n_{\rm Ecorr}$, the number of datasets for which a discovery was claimed and for which the true value of E falls within the intervals supplied, $f_{\rm Ecorr}$, the fraction of datasets for which the true value of E is in the interval, and similarly for the signal intervals, $n_{\rm Dcorr}$ and $f_{\rm Dcorr}$. The columns $\langle$Ewid$\rangle$ and $\langle$Dwid$\rangle$ indicate the average interval widths for E and D, respectively. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$.

Category	$E_{\rm true}$	$D_{\rm true}$	$n_{\rm rep}$	$n_{\rm disc}$	$f_{\rm disc}$	$n_{\rm Ecorr}$	$f_{\rm Ecorr}$	$n_{\rm Dcorr}$	$f_{\rm Dcorr}$	$\langle$Ewid$\rangle$	$\langle$Dwid$\rangle$
1	—	0.00	15400	149	0.0097 ± 0.0008	—	—	6	0.0403	0.0399	309.1451
2	0.50	83.78	200	33	0.1650 ± 0.0262	25	0.7576	26	0.7879	0.0413	163.1197
3	0.38	265.96	200	108	0.5400 ± 0.0352	82	0.7593	91	0.8426	0.0330	270.4944
4	0.10	1010.65	200	63	0.3150 ± 0.0328	57	0.9048	48	0.7619	0.0263	1000.3658
5	0.10	478.73	200	7	0.0350 ± 0.0130	5	0.7143	0	0.0000	0.0320	888.4238
6	0.66	66.49	200	39	0.1950 ± 0.0280	28	0.7179	35	0.8974	0.0406	117.1397
7	0.78	39.89	200	36	0.1800 ± 0.0272	28	0.7778	33	0.9167	0.0445	89.9875
8	0.10	744.69	200	24	0.1200 ± 0.0230	18	0.7500	14	0.5833	0.0278	914.8707
9	0.50	136.97	200	122	0.6100 ± 0.0345	102	0.8361	107	0.8770	0.0350	177.6496
10	0.90	15.29	200	13	0.0650 ± 0.0174	7	0.5385	12	0.9231	0.0524	75.2728
11	0.50	190.16	200	161	0.8050 ± 0.0280	131	0.8137	148	0.9193	0.0317	186.9565
12	0.14	664.90	200	34	0.1700 ± 0.0266	27	0.7941	23	0.6765	0.0289	790.7153
13	0.50	163.57	200	141	0.7050 ± 0.0322	106	0.7518	127	0.9007	0.0334	181.7568
14	0.38	531.92	200	169	0.8450 ± 0.0256	130	0.7692	142	0.8402	0.0214	308.6828
15	0.14	1196.83	200	96	0.4800 ± 0.0353	82	0.8542	89	0.9271	0.0240	807.8491
16	0.50	110.37	200	77	0.3850 ± 0.0344	54	0.7013	61	0.7922	0.0372	173.8191
17	0.10	1276.62	200	95	0.4750 ± 0.0353	80	0.8421	81	0.8526	0.0249	1001.8887
18	0.90	20.61	200	34	0.1700 ± 0.0266	25	0.7353	27	0.7941	0.0492	71.4330
19	0.66	132.98	200	117	0.5850 ± 0.0348	90	0.7692	108	0.9231	0.0319	135.6544
20	0.90	12.63	200	19	0.0950 ± 0.0207	12	0.6316	11	0.5789	0.0474	137.4571
21	0.90	17.95	200	27	0.1350 ± 0.0242	19	0.7037	21	0.7778	0.0491	85.2629
22	0.90	23.27	200	34	0.1700 ± 0.0266	28	0.8235	32	0.9412	0.0500	72.9724
23	0.78	79.79	200	84	0.4200 ± 0.0349	73	0.8690	75	0.8929	0.0379	106.7026
24	0.10	1542.58	200	104	0.5200 ± 0.0353	88	0.8462	91	0.8750	0.0217	1026.1458

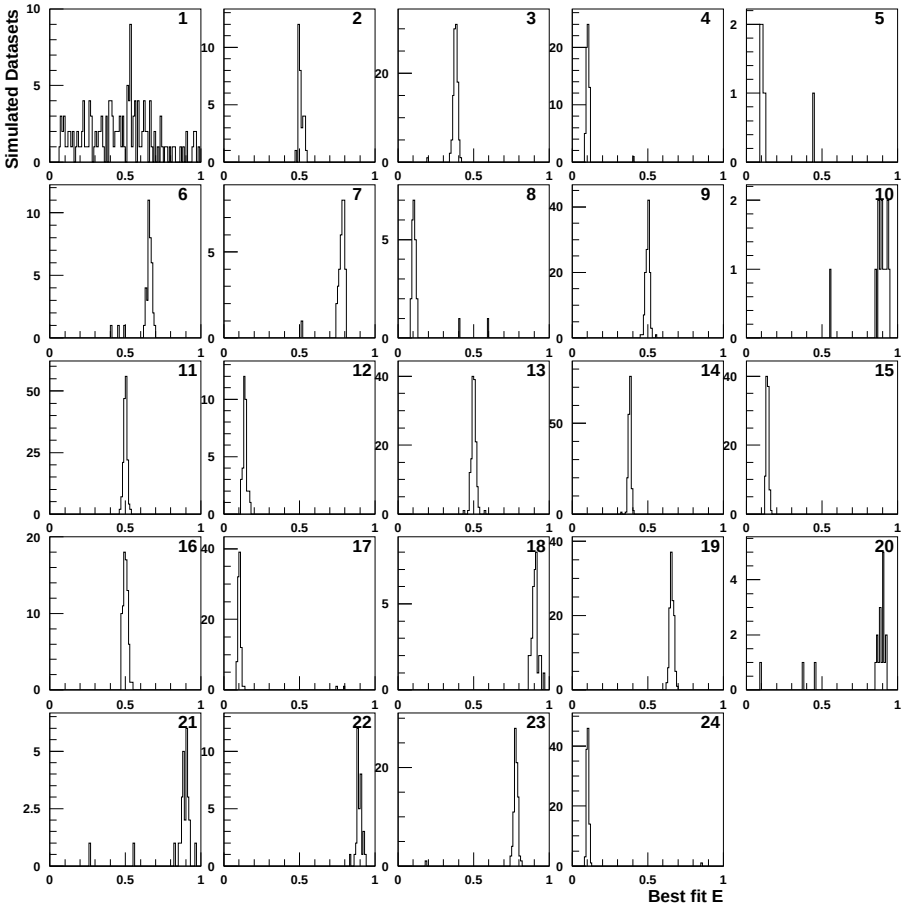

Figure 1: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Tom claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. Tom starts his fit with a guess of 0.5 explaining the peak in category 1.

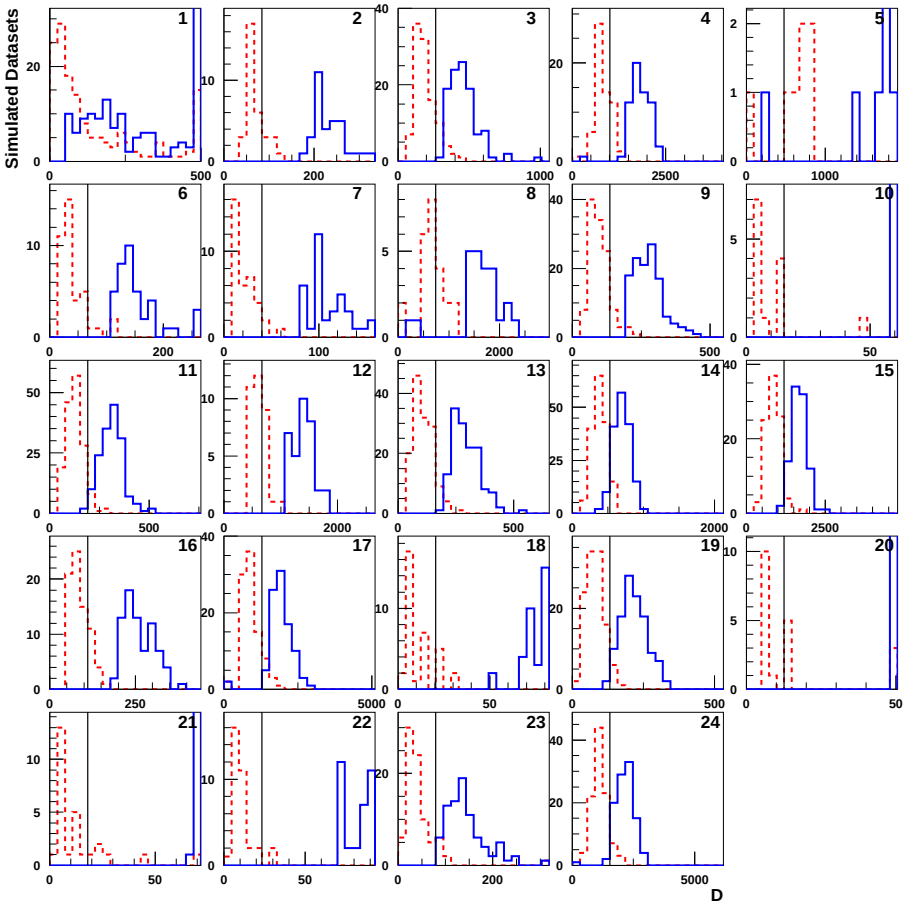

Figure 2: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Tom claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

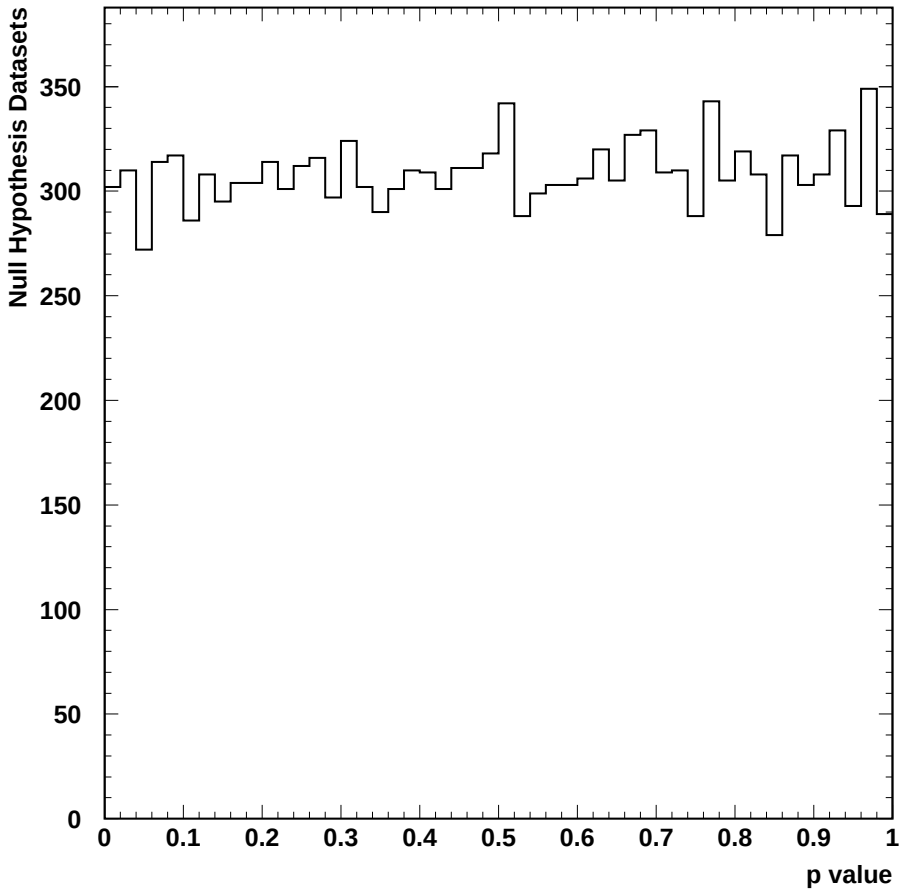

Figure 3: Distribution of the quoted p value in null hypothesis challenge datasets for Tom's solution to Problem 1.

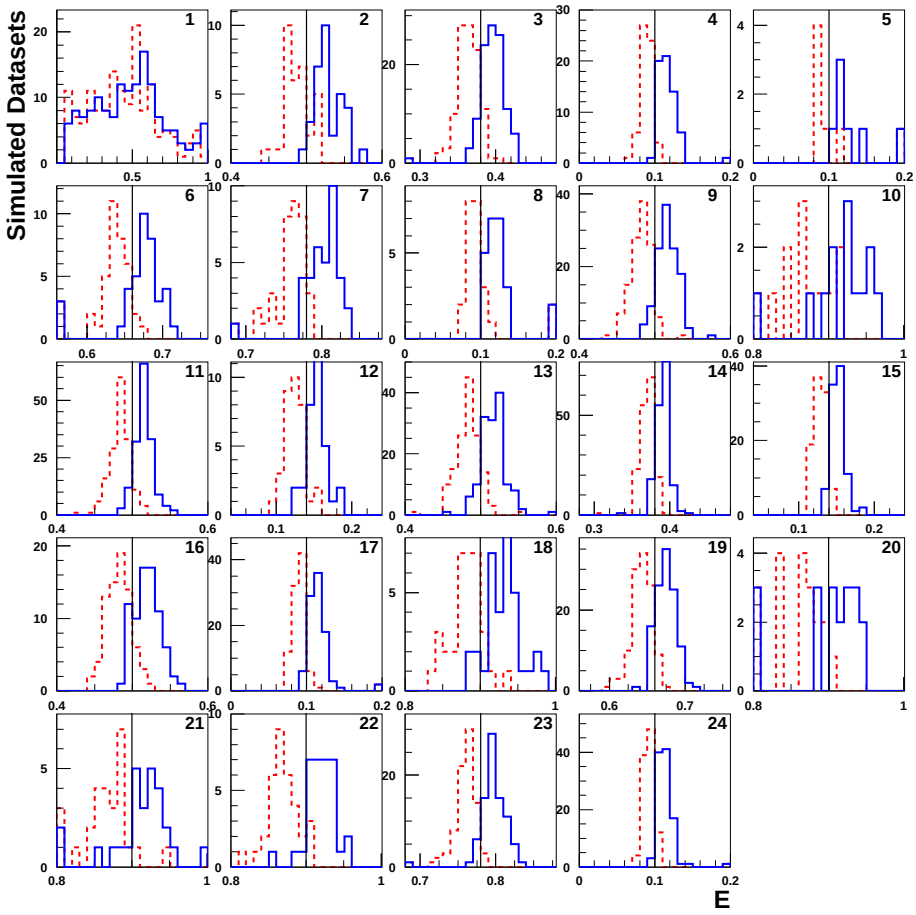

Figure 4: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Tom claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.2 From Wolfgang Rolke

For Challenge Problem #1, Wolfgang Rolke provided a solution based on a log-likelihood ratio test statistic performing two fits to each dataset. The distribution of the test statistic is predicted using simulation, since no χ^2 distribution models it for any value of the number of degrees of freedom. The critical value of the log-likelihood ratio depends on the sample size but Wolfgang found it to be roughly 11.5 for the different sample sizes in the challenge datasets. The Look-Elsewhere Effect is handled by allowing any value of E to be fit in the simulated null hypothesis datasets used to calibrate the critical value.

Table 4 lists the error rates in the challenge datasets for Wolfgang's solution. The Type-I error rate is 1% as desired.

Table 4: Problem 1 performance evaluation for Wolfgang Rolke's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	$E_{\rm true}$	$D_{\rm true}$	$n_{\rm rep}$	$n_{\rm disc}$	$f_{\rm disc}$	$n_{\rm Ecorr}$	$f_{\rm Ecorr}$	$n_{\rm Dcorr}$	$f_{\rm Dcorr}$	$\langle$Ewid$\rangle$	$\langle$Dwid$\rangle$
1	—	0.00	15400	159	0.0103 ± 0.0008	—	—	0	0.0000	0.0249	398.0893
2	0.50	83.78	200	23	0.1150 ± 0.0226	13	0.5652	1	0.0435	0.0262	116.7739
3	0.38	265.96	200	112	0.5600 ± 0.0351	66	0.5893	70	0.6250	0.0236	193.4187
4	0.10	1010.65	200	76	0.3800 ± 0.0343	51	0.6711	33	0.4342	0.0196	712.8289
5	0.10	478.73	200	5	0.0250 ± 0.0110	2	0.4000	0	0.0000	0.0236	737.1400
6	0.66	66.49	200	64	0.3200 ± 0.0330	39	0.6094	43	0.6719	0.0268	79.8875
7	0.78	39.89	200	72	0.3600 ± 0.0339	43	0.5972	45	0.6250	0.0281	70.1986
8	0.10	744.69	200	40	0.2000 ± 0.0283	18	0.4500	3	0.0750	0.0198	685.9375
9	0.50	136.97	200	105	0.5250 ± 0.0353	68	0.6476	57	0.5429	0.0248	126.5714
10	0.90	15.29	200	15	0.0750 ± 0.0186	6	0.4000	9	0.6000	0.0311	52.1133
11	0.50	190.16	200	158	0.7900 ± 0.0288	105	0.6646	122	0.7722	0.0222	131.1456
12	0.14	664.90	200	49	0.2450 ± 0.0304	30	0.6122	9	0.1837	0.0221	547.3857
13	0.50	163.57	200	126	0.6300 ± 0.0341	78	0.6190	81	0.6429	0.0230	128.4675
14	0.38	531.92	200	199	0.9950 ± 0.0050	139	0.6985	145	0.7286	0.0153	218.2327
15	0.14	1196.83	200	185	0.9250 ± 0.0186	127	0.6865	140	0.7568	0.0174	573.9854
16	0.50	110.37	200	66	0.3300 ± 0.0332	31	0.4697	23	0.3485	0.0273	123.5334
17	0.10	1276.62	200	135	0.6750 ± 0.0331	94	0.6963	87	0.6444	0.0180	732.7238
18	0.90	20.61	200	53	0.2650 ± 0.0312	31	0.5849	23	0.4340	0.0310	63.5094
19	0.66	132.98	200	184	0.9200 ± 0.0192	116	0.6304	139	0.7554	0.0228	93.8962
20	0.90	12.63	200	27	0.1350 ± 0.0242	15	0.5556	11	0.4074	0.0305	78.9630
21	0.90	17.95	200	43	0.2150 ± 0.0290	24	0.5581	19	0.4419	0.0310	70.2302
22	0.90	23.27	200	52	0.2600 ± 0.0310	31	0.5962	26	0.5000	0.0322	55.3750
23	0.78	79.79	200	161	0.8050 ± 0.0280	111	0.6894	130	0.8075	0.0259	71.6441
24	0.10	1542.58	200	178	0.8900 ± 0.0221	130	0.7303	132	0.7416	0.0158	754.2759

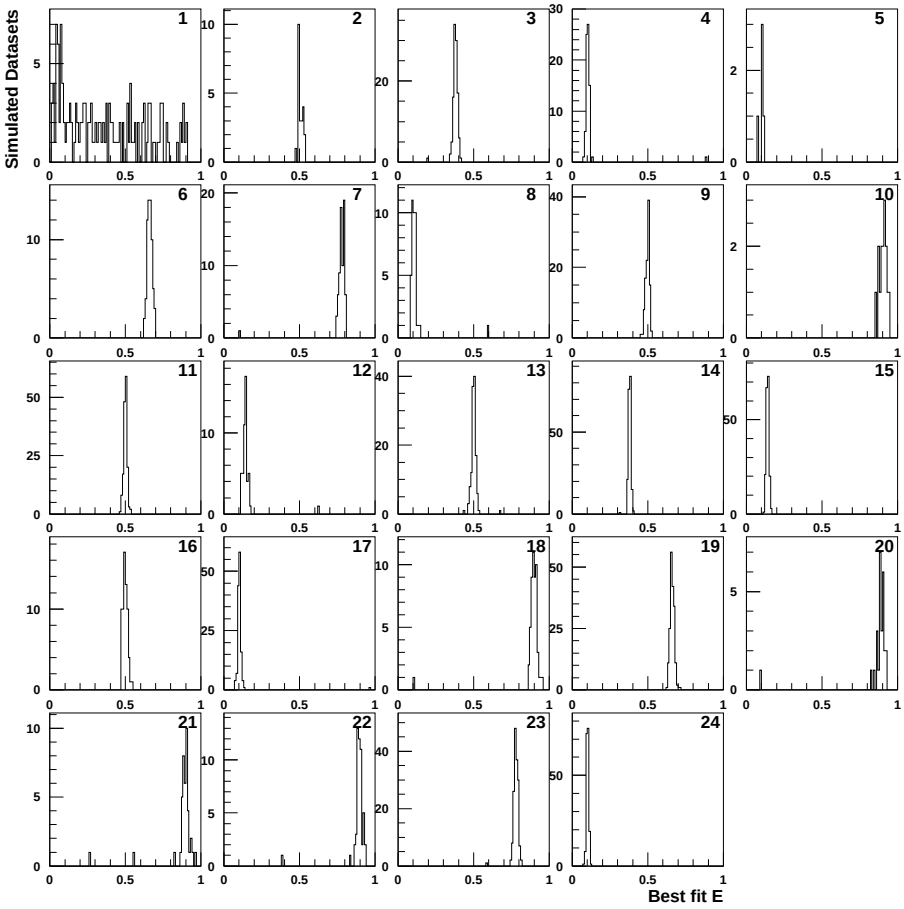

Figure 5: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Wolfgang claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

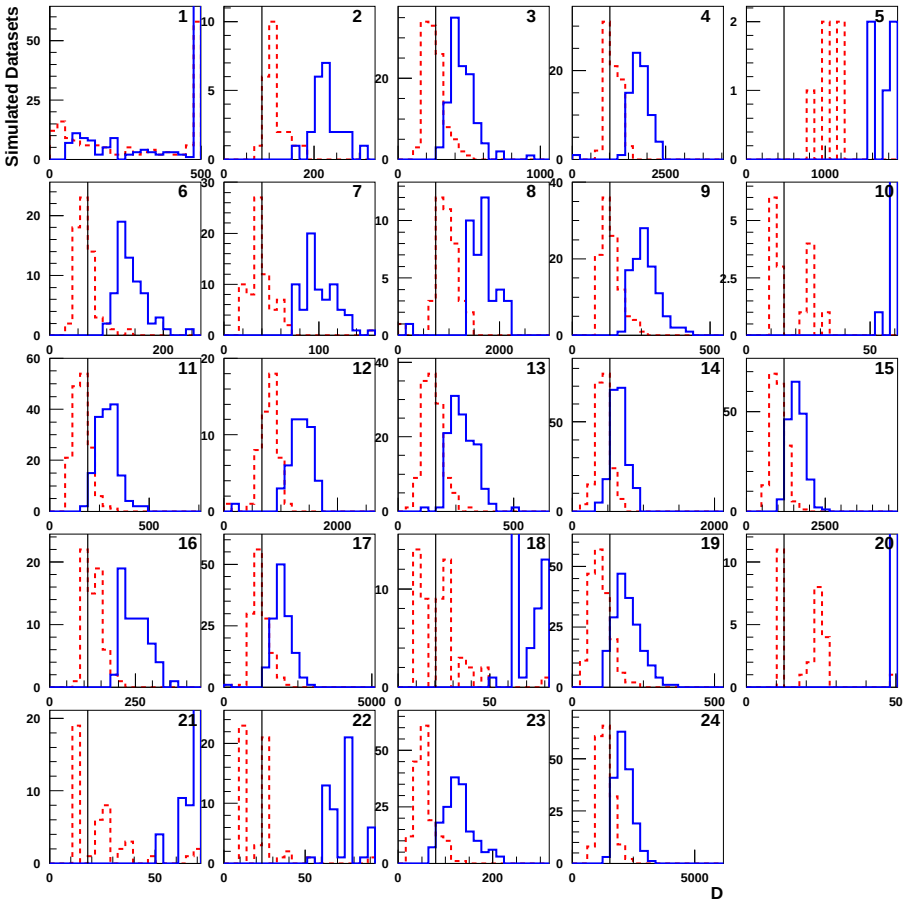

Figure 6: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Wolfgang claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

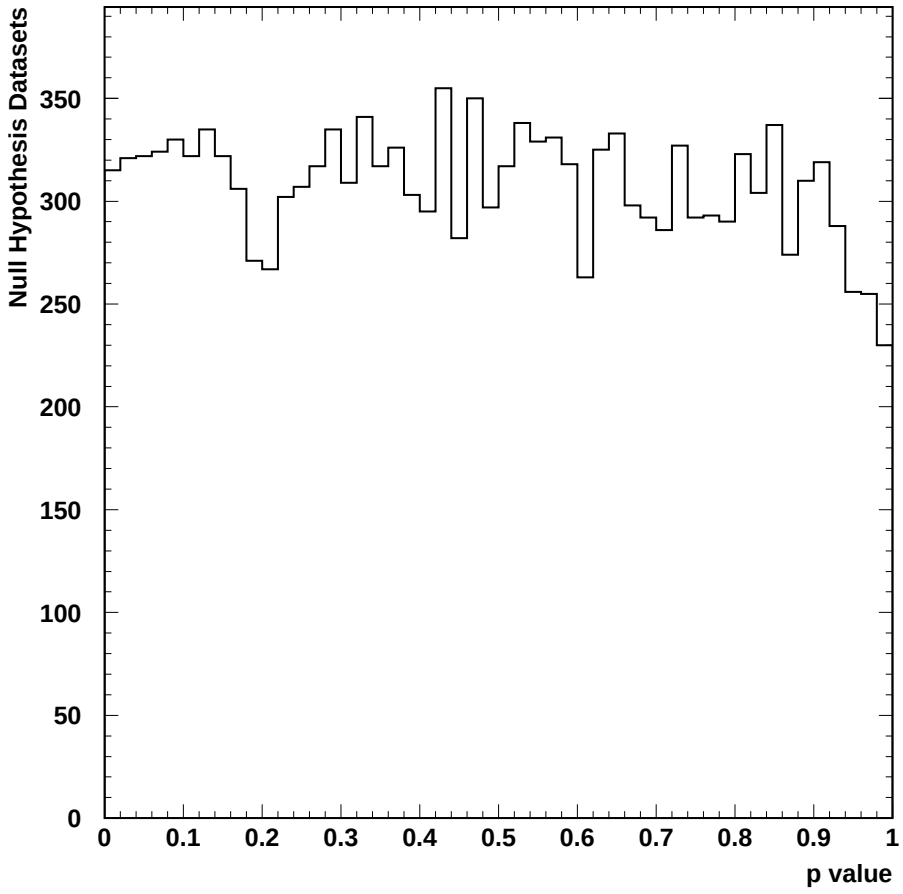

Figure 7: Distribution of the quoted p value in null hypothesis challenge datasets for Wolfgang's solution to Problem 1.

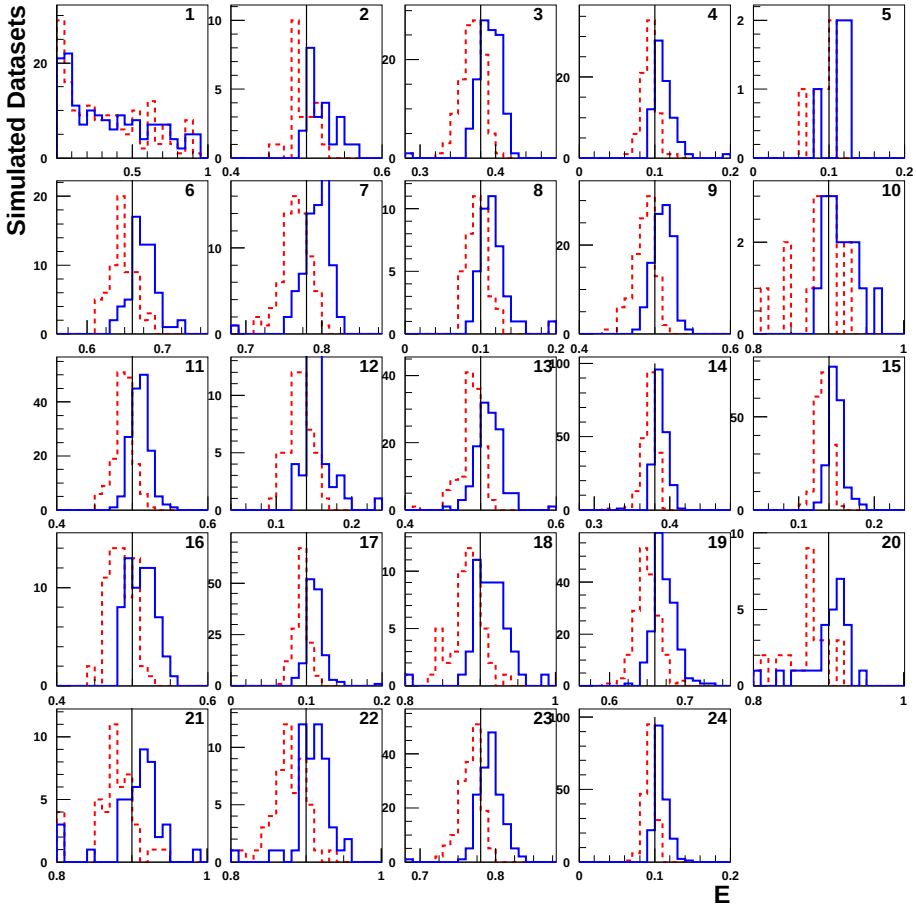

Figure 8: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Wolfgang claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.3 From the Stanford Challenge Team

The Stanford Challenge Team (SCT) consists of Brad Efron, Trevor Hastie, Omkar Muralidharan, Balasubramanian Narasimhan, Jeffrey Scargle, Rob Tibshirani, and Ryan Tibshirani. The SCT provided a solution to Problem 1 based on a log-likelihood ratio test statistic performing two fits to each dataset. The distribution of the test statistic is predicted using simulation. The Look-Elsewhere Effect is handled by allowing any value of E to be fit in the simulated null hypothesis datasets used to calibrate the critical value. The parameters D and E were fit for using a maximum-likelihood approach, and used the non-parametric bootstrap to estimate the variability of the results.

Table 5 lists the error rates in the challenge datasets for the SCT's solution. The Type-I error rate is just under 1% as desired. The upturn in the distribution of the quoted p values for null outcomes shown in Figure 11 at high quoted p values is an indication of the slight overcoverage at small values of the quoted p value.

Table 5: Problem 1 performance evaluation for the Stanford Challenge Team's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	$\langle$Ewid$\rangle$	$\langle$Dwid$\rangle$
1	—	0.00	15400	118	0.0077 ± 0.0007	—	—	0	0.0000	0.0385	242.6723
2	0.50	83.78	200	23	0.1150 ± 0.0226	17	0.7391	6	0.2609	0.0448	115.4431
3	0.38	265.96	200	112	0.5600 ± 0.0351	80	0.7143	74	0.6607	0.0331	193.5983
4	0.10	1010.65	200	71	0.3550 ± 0.0338	50	0.7042	39	0.5493	0.0216	709.0595
5	0.10	478.73	200	5	0.0250 ± 0.0110	3	0.6000	0	0.0000	0.0263	730.5103
6	0.66	66.49	200	64	0.3200 ± 0.0330	43	0.6719	47	0.7344	0.0342	76.7652
7	0.78	39.89	200	71	0.3550 ± 0.0338	49	0.6901	31	0.4366	0.0359	65.9234
8	0.10	744.69	200	38	0.1900 ± 0.0277	23	0.6053	5	0.1316	0.0244	691.4062
9	0.50	136.97	200	104	0.5200 ± 0.0353	77	0.7404	61	0.5865	0.0320	126.3906
10	0.90	15.29	200	15	0.0750 ± 0.0186	12	0.8000	4	0.2667	0.0589	42.1555
11	0.50	190.16	200	158	0.7900 ± 0.0288	116	0.7342	125	0.7911	0.0281	130.5645
12	0.14	664.90	200	49	0.2450 ± 0.0304	30	0.6122	12	0.2449	0.0263	545.1991
13	0.50	163.57	200	126	0.6300 ± 0.0341	86	0.6825	83	0.6587	0.0300	128.7363
14	0.38	531.92	200	199	0.9950 ± 0.0050	134	0.6734	144	0.7236	0.0173	219.5495
15	0.14	1196.83	200	185	0.9250 ± 0.0186	119	0.6432	146	0.7892	0.0206	577.7383
16	0.50	110.37	200	65	0.3250 ± 0.0331	40	0.6154	29	0.4462	0.0379	122.6103
17	0.10	1276.62	200	135	0.6750 ± 0.0331	104	0.7704	95	0.7037	0.0204	720.1168
18	0.90	20.61	200	53	0.2650 ± 0.0312	45	0.8491	25	0.4717	0.0541	53.5806
19	0.66	132.98	200	184	0.9200 ± 0.0192	109	0.5924	130	0.7065	0.0255	91.4350
20	0.90	12.63	200	27	0.1350 ± 0.0242	20	0.7407	0	0.0000	0.0562	69.1609
21	0.90	17.95	200	42	0.2100 ± 0.0288	31	0.7381	7	0.1667	0.0578	48.4375
22	0.90	23.27	200	52	0.2600 ± 0.0310	44	0.8462	33	0.6346	0.0544	45.1804
23	0.78	79.79	200	159	0.7950 ± 0.0285	114	0.7170	114	0.7170	0.0321	65.9111
24	0.10	1542.58	200	177	0.8850 ± 0.0226	141	0.7966	139	0.7853	0.0175	750.7211

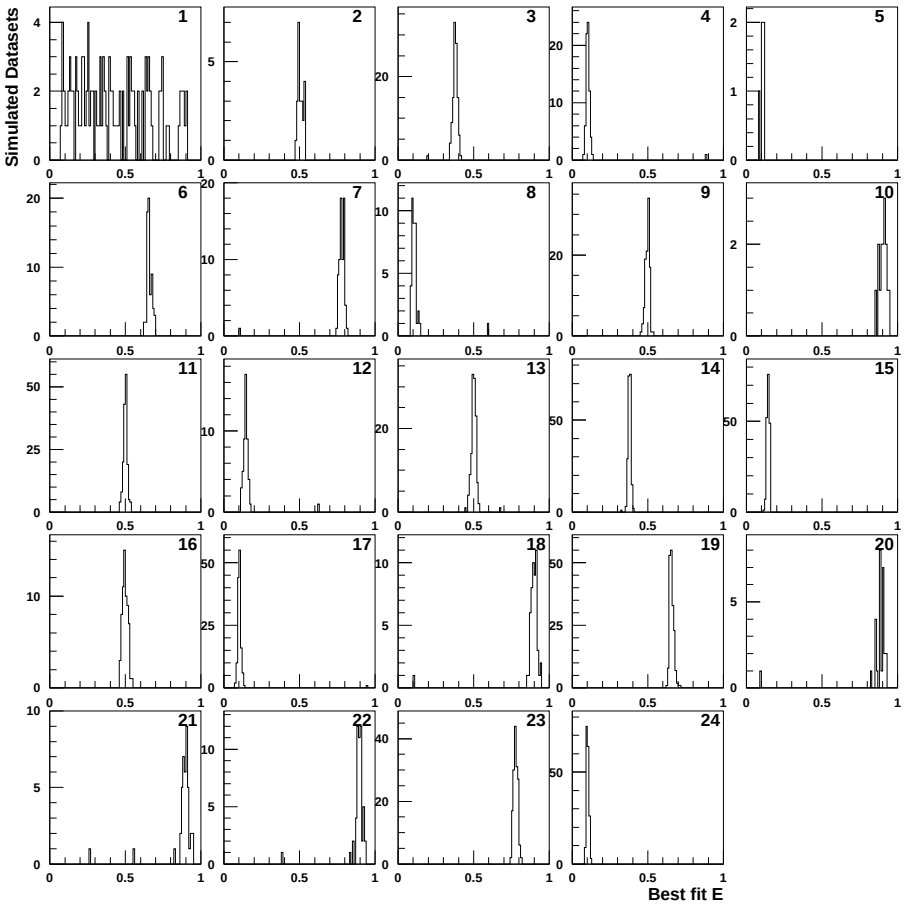

Figure 9: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which the SCT claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

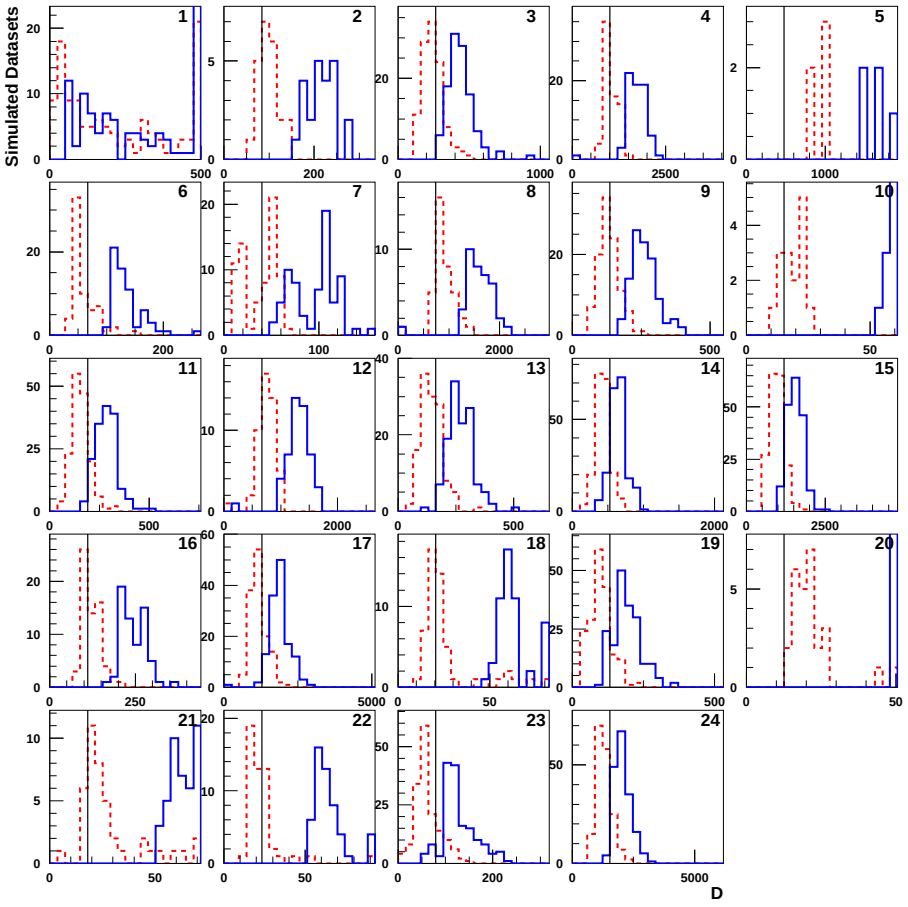

Figure 10: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which the SCT claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

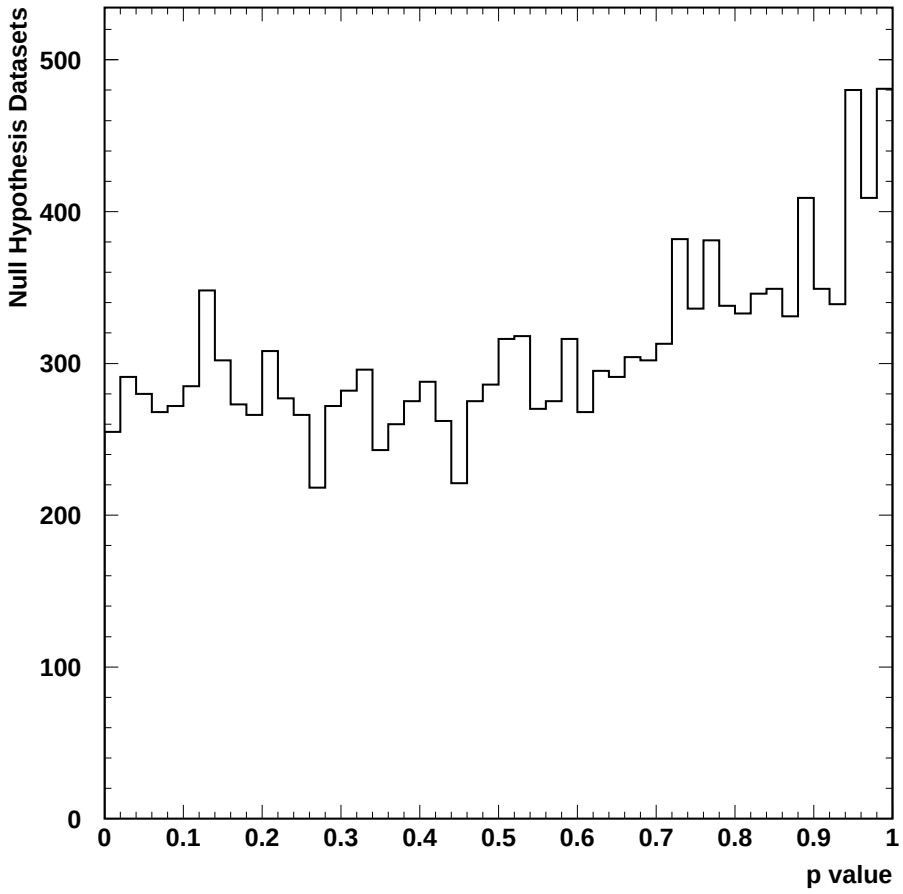

Figure 11: Distribution of the quoted p value in null hypothesis challenge datasets for the SCT's solution to Problem 1.

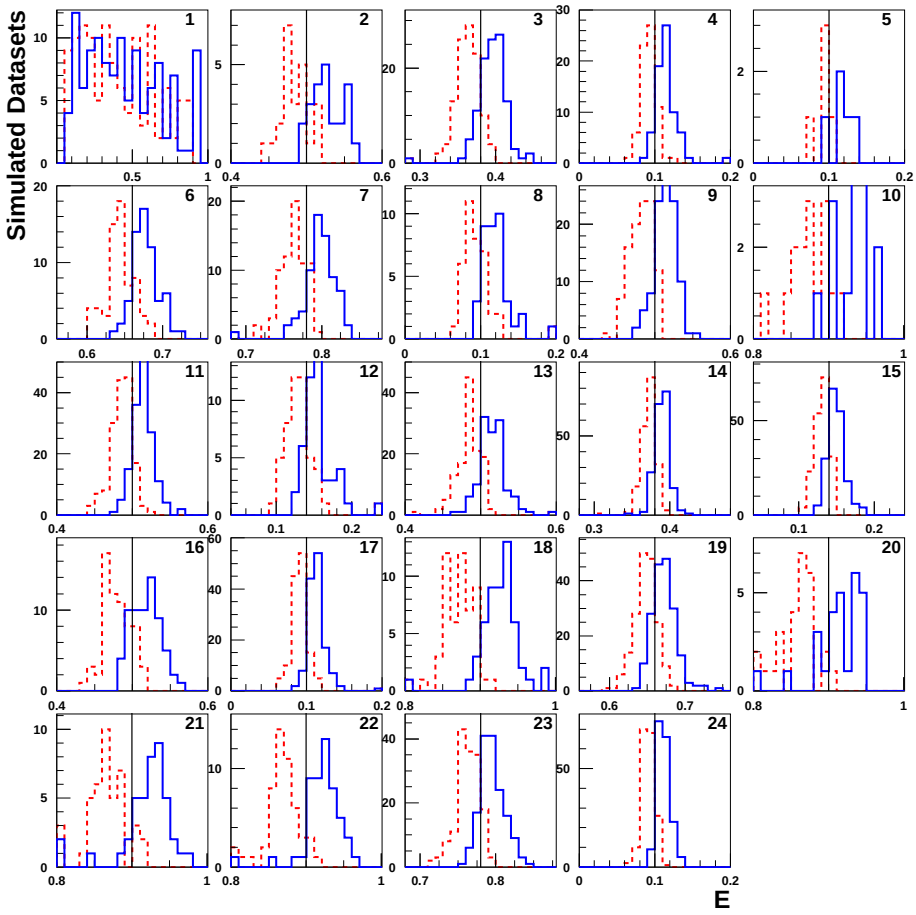

Figure 12: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which the SCT claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.4 From Eilam Gross and Ofer Vitells

Eilam and Ofer provided a solution to Problem 1 based on a two-fit log likelihood ratio similar to those used by other participants. The Look-Elsewhere Effect is addressed using a procedure described in [1]. Confidence intervals for D and E are computed using $\Delta 2 \log \lambda = 1$, additionally setting the lower bound on the signal rate to be zero when $P(q_0 \leq q_0^{observed} | H_0) = 68\%$.

Table 6 lists the error rates in the challenge datasets for Eilam and Ofer's solution. The Type-I error rate is just under 1% as desired. The distribution of the p values in the null datasets, shown in Figure 15 is interesting, as the quoted p values exceed unity. This is presumably a consequence of the procedure used to account for the trials factor. Since it affects large p values and not the ones near the critical point of 0.01, it does not have an impact on the result. Typically if there is insufficient evidence for a signal, a particle physics experiment does not compute a p value and instead quotes upper limits on the signal rate, and so the large p values do not have an impact on any results.

Table 6: Problem 1 performance evaluation for Eilam and Ofer's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	$E_{\rm true}$	$D_{\rm true}$	$n_{\rm rep}$	$n_{\rm disc}$	$f_{\rm disc}$	$n_{\rm Ecorr}$	$f_{\rm Ecorr}$	$n_{\rm Dcorr}$	$f_{\rm Dcorr}$	⟨Ewid⟩	⟨Dwid⟩
1	—	0.00	15400	124	0.0081 ± 0.0007	—	—	0	0.0000	0.0274	242.2684
2	0.50	83.78	200	24	0.1200 ± 0.0230	16	0.6667	2	0.0833	0.0291	118.6147
3	0.38	265.96	200	114	0.5700 ± 0.0350	68	0.5965	74	0.6491	0.0227	195.0962
4	0.10	1010.65	200	72	0.3600 ± 0.0339	44	0.6111	36	0.5000	0.0177	725.0212
5	0.10	478.73	200	5	0.0250 ± 0.0110	2	0.4000	0	0.0000	0.0190	741.5024
6	0.66	66.49	200	64	0.3200 ± 0.0330	38	0.5938	42	0.6562	0.0262	80.9731
7	0.78	39.89	200	74	0.3700 ± 0.0341	48	0.6486	40	0.5405	0.0299	72.7214
8	0.10	744.69	200	39	0.1950 ± 0.0280	17	0.4359	4	0.1026	0.0187	695.0071
9	0.50	136.97	200	105	0.5250 ± 0.0353	64	0.6095	61	0.5810	0.0242	128.7545
10	0.90	15.29	200	16	0.0800 ± 0.0192	7	0.4375	0	0.0000	0.0361	49.7170
11	0.50	190.16	200	160	0.8000 ± 0.0283	108	0.6750	130	0.8125	0.0215	133.4258
12	0.14	664.90	200	49	0.2450 ± 0.0304	27	0.5510	12	0.2449	0.0202	559.3315
13	0.50	163.57	200	126	0.6300 ± 0.0341	77	0.6111	83	0.6587	0.0242	130.6738
14	0.38	531.92	200	199	0.9950 ± 0.0050	111	0.5578	149	0.7487	0.0133	221.4849
15	0.14	1196.83	200	185	0.9250 ± 0.0186	126	0.6811	144	0.7784	0.0168	583.2854
16	0.50	110.37	200	67	0.3350 ± 0.0334	36	0.5373	27	0.4030	0.0260	125.5202
17	0.10	1276.62	200	135	0.6750 ± 0.0331	89	0.6593	95	0.7037	0.0162	736.7133
18	0.90	20.61	200	54	0.2700 ± 0.0314	35	0.6481	31	0.5741	0.0333	62.8133
19	0.66	132.98	200	184	0.9200 ± 0.0192	72	0.3913	147	0.7989	0.0167	94.8142
20	0.90	12.63	200	27	0.1350 ± 0.0242	18	0.6667	0	0.0000	0.0351	77.9128
21	0.90	17.95	200	42	0.2100 ± 0.0288	25	0.5952	0	0.0000	0.0322	57.2445
22	0.90	23.27	200	52	0.2600 ± 0.0310	35	0.6731	24	0.4615	0.0330	53.6616
23	0.78	79.79	200	163	0.8150 ± 0.0275	74	0.4540	133	0.8160	0.0198	74.3283
24	0.10	1542.58	200	178	0.8900 ± 0.0221	127	0.7135	136	0.7640	0.0151	765.4595

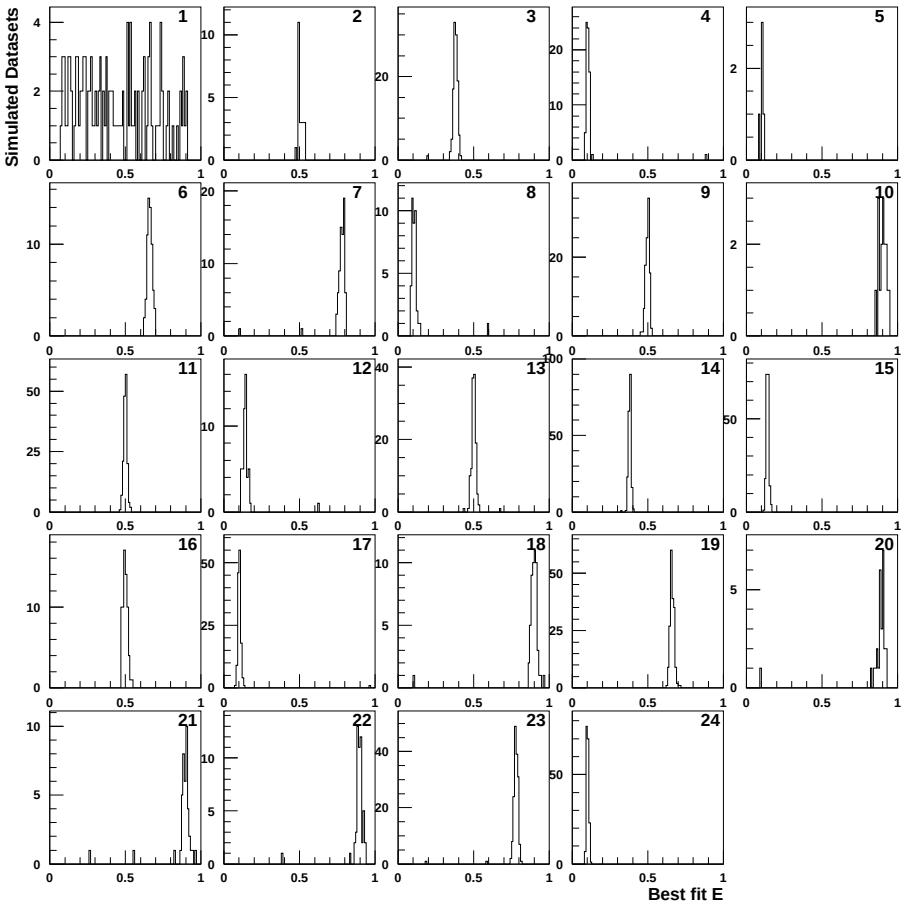

Figure 13: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Eilam and Ofer claim evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

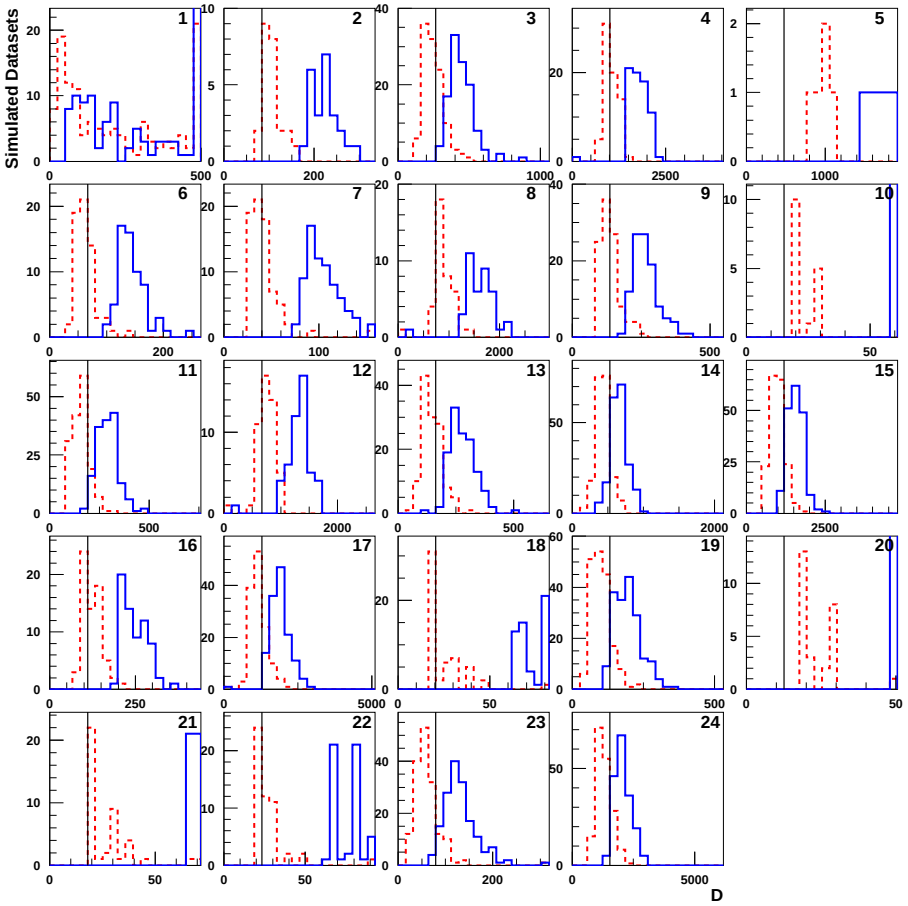

Figure 14: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Eilam and Ofer claim evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

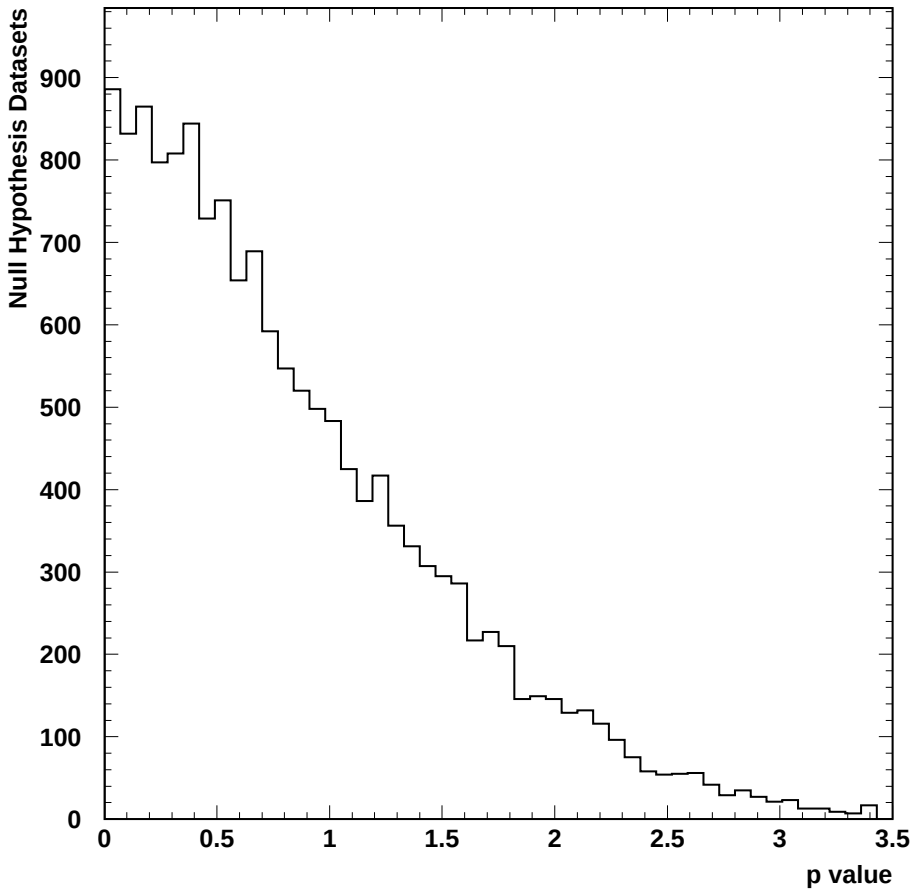

Figure 15: Distribution of the quoted p value in null hypothesis challenge datasets for Eilam and Ofer's solution to Problem 1. The look-elsewhere correction factor makes the maximum value of the p value exceed 1.0.

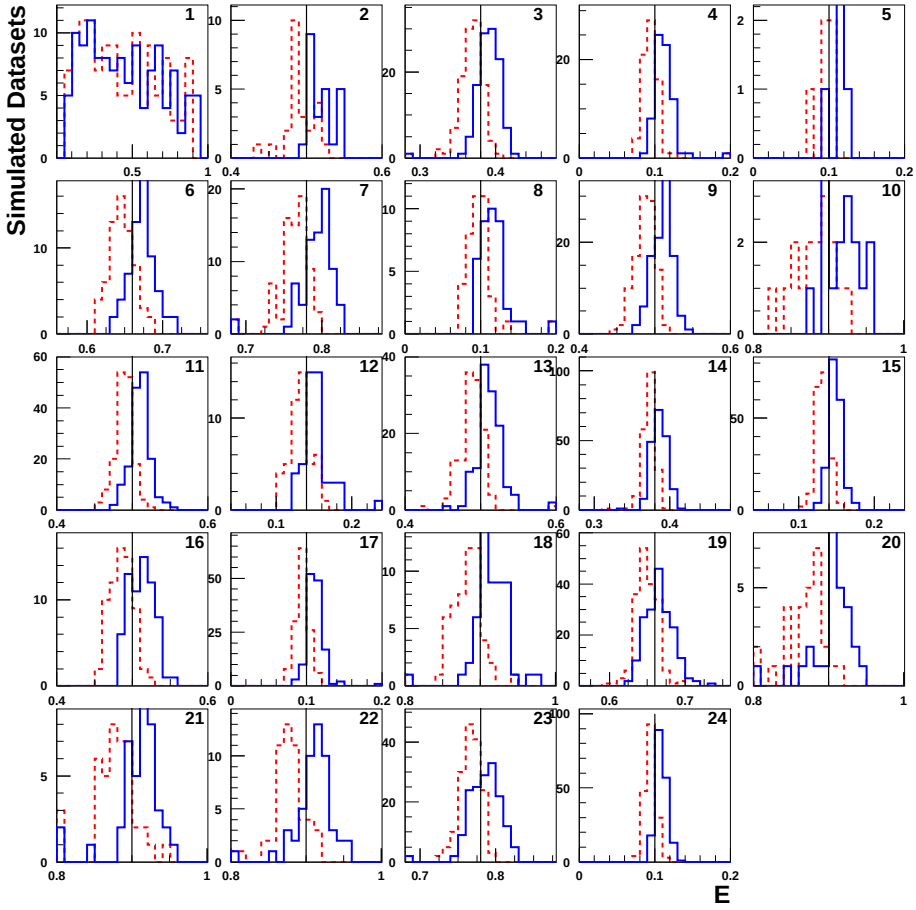

Figure 16: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Eilam and Ofer claim evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.5 From Valentin Niess

For Challenge Problem #1, Valentin Niess provided a solution based on a method which counts events within a bracketing interval chosen to maximize the sensitivity to find a signal. The half width of the bracketing interval is 1.4σ where $\sigma = 0.03$, the signal width in E. A search over N_{bin} brackets allows the best-fit value of E to be anywhere in the range $0 \leq E \leq 1$. An effective dimension $N_{eff} = 9$ provides a factor to adjust the p value for the Look-Elsewhere Effect.

Table 7 lists the error rates in the challenge datasets for Valentin's solution. The Type-I error rate is not measurably different from 1% as desired. Even though Valentin's p value distribution tilts up at large values of the p value, as can be seen in Figure 19, the distribution rises again for $p < 0.01$, giving approximately the correct coverage at this choice of the desired Type-I error rate. Valentin's estimate Type-II error rates, listed in Table 2 were computed using a sample of simulated datasets with a fixed number of marks in them – 1000. Since the sample of null-hypothesis simulated datasets in the Challenge sample was drawn from a different sample space – varying A and choosing Poisson-fluctuated data from the varied value of A for each simulated dataset, we do not expect the error rates to match up perfectly.

Valentin recomputed his expected sensitivities releasing the total event count of 1000 requirement, which had the side effect of reducing the background rate for large signals and increasing it for small signals. The new powers are:

$$\begin{array}{rrr} D = 1010 & E = 0.1 & \beta = 0.34 \\ D = 137 & E = 0.5 & \beta = 0.46 \\ D = 18 & E = 0.9 & \beta = 0.17 \end{array}$$

Since these came in after the solutions were released, the tables and plots remain the same.

Table 7: Problem 1 performance evaluation for Valentin Niess's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	$\langle$Ewid$\rangle$	$\langle$Dwid$\rangle$
1	—	0.00	15400	171	0.0111 ± 0.0008	—	—	0	0.0000	0.0365	436.3579
2	0.50	83.78	200	19	0.0950 ± 0.0207	14	0.7368	0	0.0000	0.0369	76.5268
3	0.38	265.96	200	114	0.5700 ± 0.0350	82	0.7193	76	0.6667	0.0340	148.9114
4	0.10	1010.65	200	65	0.3250 ± 0.0331	51	0.7846	37	0.5692	0.0310	772.5247
5	0.10	478.73	200	6	0.0300 ± 0.0121	5	0.8333	0	0.0000	0.0432	690.0167
6	0.66	66.49	200	61	0.3050 ± 0.0326	48	0.7869	31	0.5082	0.0395	33.6639
7	0.78	39.89	200	66	0.3300 ± 0.0332	47	0.7121	5	0.0758	0.0415	12.5427
8	0.10	744.69	200	30	0.1500 ± 0.0252	12	0.4000	8	0.2667	0.0270	738.9333
9	0.50	136.97	200	106	0.5300 ± 0.0353	77	0.7264	61	0.5755	0.0371	78.9471
10	0.90	15.29	200	17	0.0850 ± 0.0197	8	0.4706	1	0.0588	0.0432	10.2747
11	0.50	190.16	200	156	0.7800 ± 0.0293	122	0.7821	97	0.6218	0.0349	78.0085
12	0.14	664.90	200	44	0.2200 ± 0.0293	31	0.7045	20	0.4545	0.0358	563.5585
13	0.50	163.57	200	127	0.6350 ± 0.0340	97	0.7638	81	0.6378	0.0367	78.8177
14	0.38	531.92	200	197	0.9850 ± 0.0086	149	0.7563	87	0.4416	0.0273	146.4340
15	0.14	1196.83	200	178	0.8900 ± 0.0221	138	0.7753	129	0.7247	0.0285	583.5477
16	0.50	110.37	200	59	0.2950 ± 0.0322	43	0.7288	18	0.3051	0.0400	77.6622
17	0.10	1276.62	200	124	0.6200 ± 0.0343	86	0.6935	92	0.7419	0.0274	787.9966
18	0.90	20.61	200	51	0.2550 ± 0.0308	32	0.6275	6	0.1176	0.0432	20.4871
19	0.66	132.98	200	182	0.9100 ± 0.0202	133	0.7308	64	0.3516	0.0353	36.9306
20	0.90	12.63	200	23	0.1150 ± 0.0226	14	0.6087	4	0.1739	0.0389	42.3548
21	0.90	17.95	200	39	0.1950 ± 0.0280	24	0.6154	3	0.0769	0.0460	19.2436
22	0.90	23.27	200	47	0.2350 ± 0.0300	29	0.6170	2	0.0426	0.0440	7.9202
23	0.78	79.79	200	158	0.7900 ± 0.0288	114	0.7215	40	0.2532	0.0382	19.4998
24	0.10	1542.58	200	165	0.8250 ± 0.0269	124	0.7515	118	0.7152	0.0276	797.7090

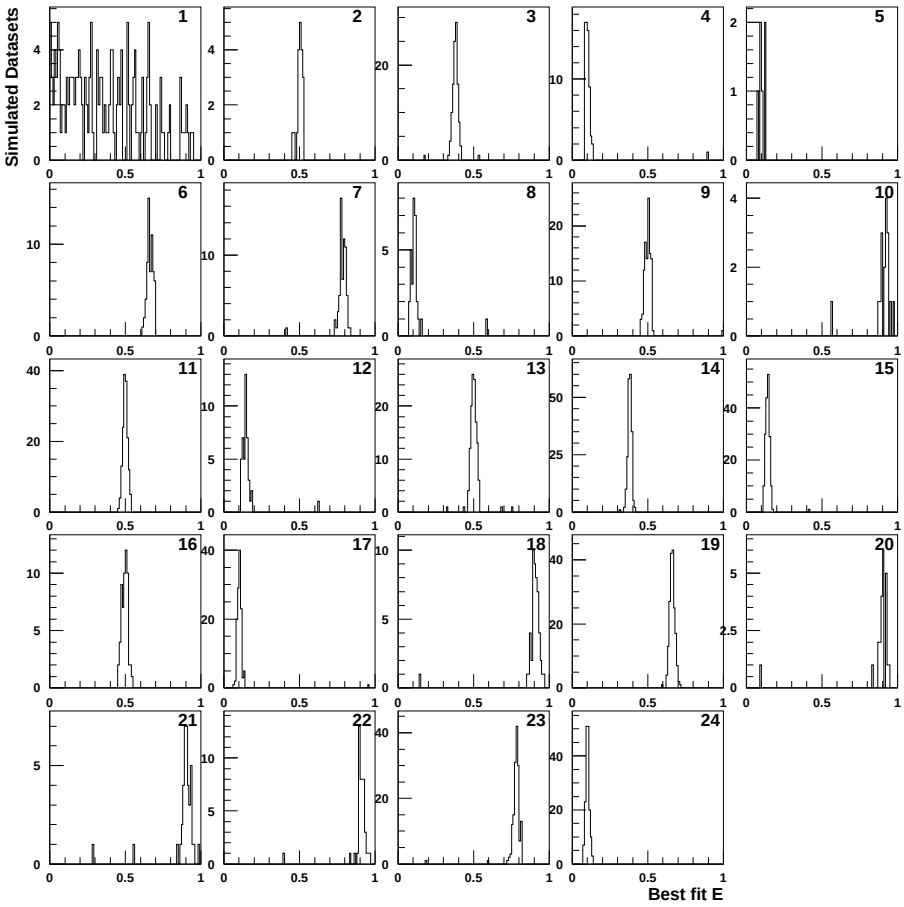

Figure 17: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Valentin claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

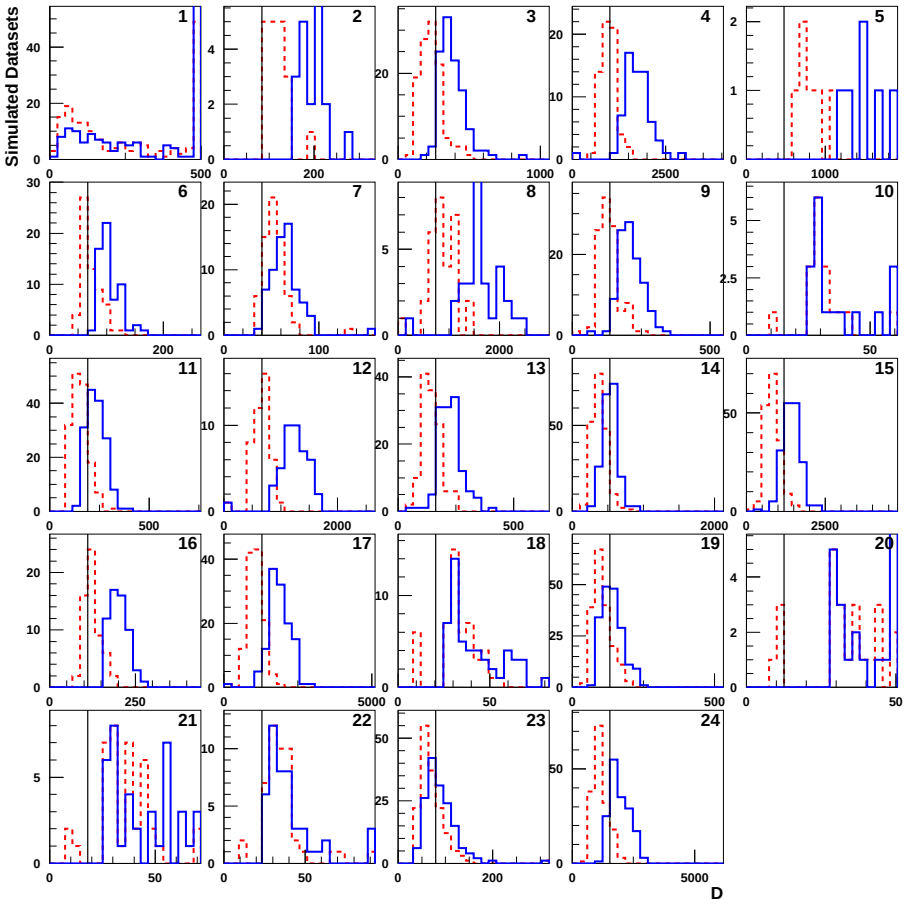

Figure 18: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Valentin claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

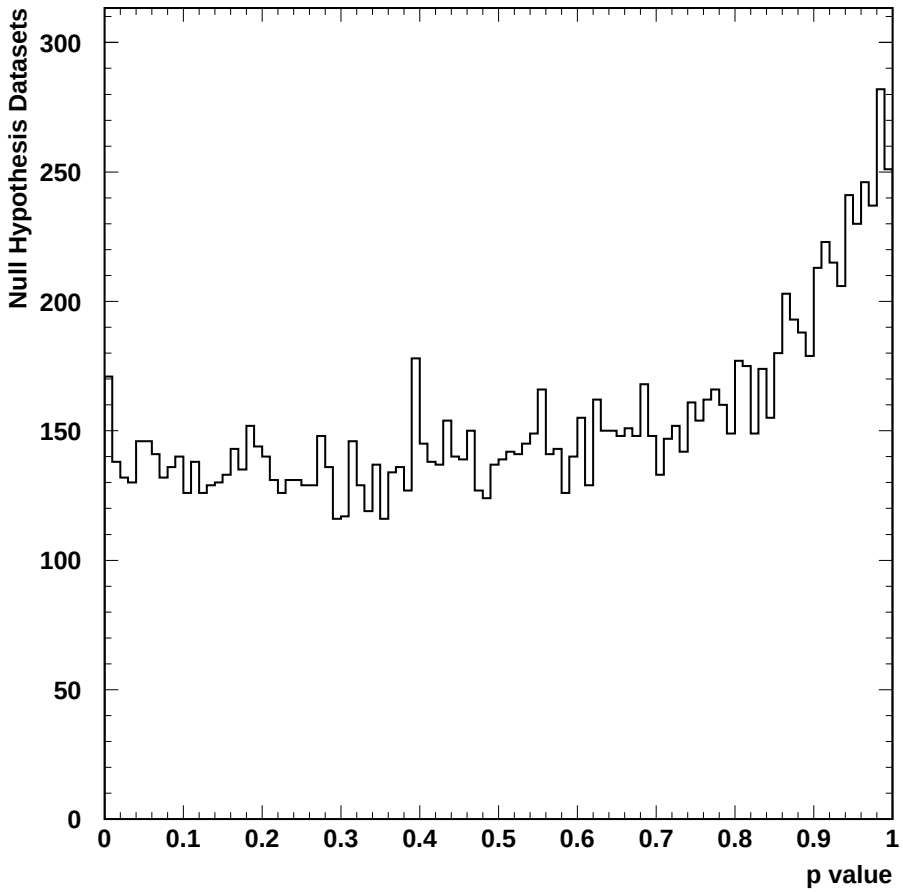

Figure 19: Distribution of the quoted p value in null hypothesis challenge datasets for Valentin's solution to Problem 1. One hundred bins are chosen for this figure to show the rise in the distribution below $p = 0.01$.

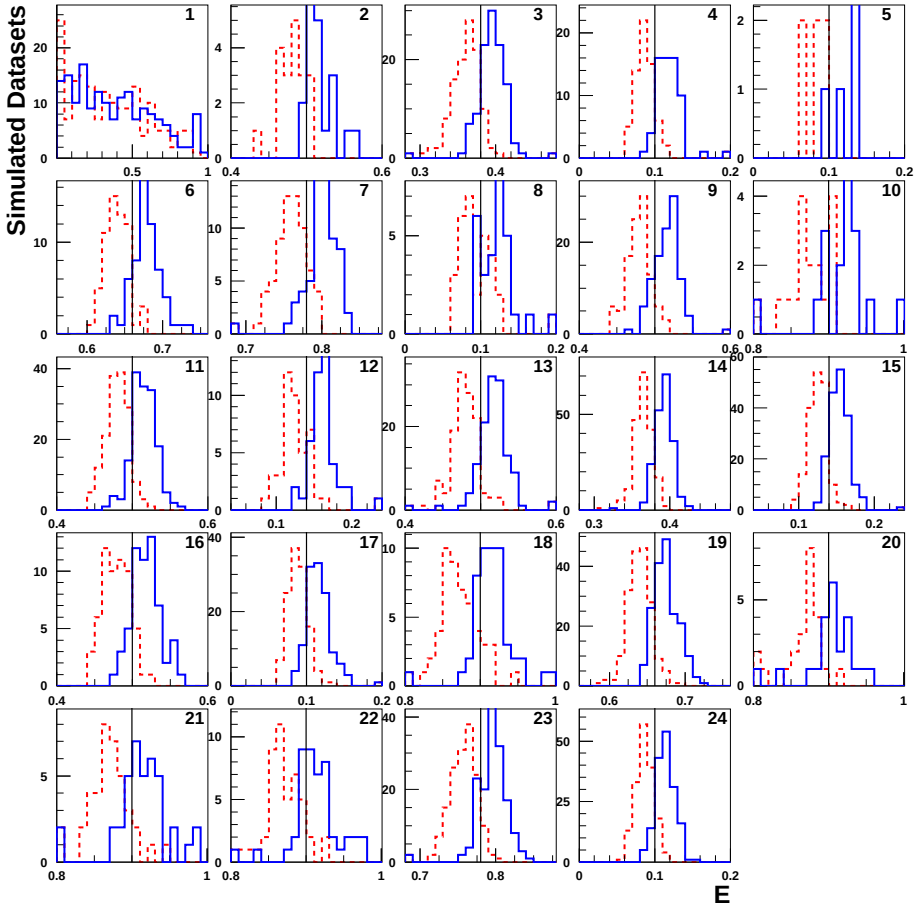

Figure 20: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Valentin claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.6 From Georgios Choudalakis

For Challenge Problem #1, Georgios Choudalakis provided a solution using his BumpHunter program [2]. In this case, the marks were binned in a histogram with 40 bins from 0 to 1. The contents of bins are collected into a signal window and sidebands on either side of the proposed signal window. The event count in the signal window and a fit to the background outside of the signal window are used to construct the test statistic for discovery. The Look-Elsewhere Effect is taken into account by constructing "hyper-tests" –

Table 8 lists the error rates in the challenge datasets for Georgios's solution. The Type-I error rate is not measurably different from 1%.

BumpHunter optimizes its computation by running simulations until there is sufficient confidence that $p > 0.01$ or that $p < 0.01$. Often, 10 simulated null repetitions are sufficient to make a decision if it is clear that a signal is not present, and sometimes it takes more. This computational optimization makes the quoted p value distribution rather discrete, particularly at large p values, as can be seen in Figure 23. More simulated null datasets are generated for smaller p values; for p values very close to 0.01, a larger amount of CPU is required to make a decision.

Table 8: Problem 1 performance evaluation for Georgios Choudalakis's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. One of the challenge datasets in signal category 9 resulted in an unusually large interval for D (perhaps a fit failure), making the average width very large. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	⟨Ewid⟩	⟨Dwid⟩
1	—	0.00	15400	169	0.0110 ± 0.0008	—	—	24	0.1420	0.0386	371.8143
2	0.50	83.78	200	19	0.0950 ± 0.0207	8	0.4211	8	0.4211	0.0270	118.8511
3	0.38	265.96	200	66	0.3300 ± 0.0332	32	0.4848	33	0.5000	0.0216	207.7717
4	0.10	1010.65	200	32	0.1600 ± 0.0259	19	0.5938	10	0.3125	0.0169	743.2350
5	0.10	478.73	200	1	0.0050 ± 0.0050	0	0.0000	0	0.0000	0.0161	662.3292
6	0.66	66.49	200	48	0.2400 ± 0.0302	30	0.6250	22	0.4583	0.0404	128.4285
7	0.78	39.89	200	64	0.3200 ± 0.0330	43	0.6719	25	0.3906	0.0542	175.9160
8	0.10	744.69	200	23	0.1150 ± 0.0226	10	0.4348	4	0.1739	0.0213	680.9976
9	0.50	136.97	200	70	0.3500 ± 0.0337	38	0.5429	40	0.5714	0.0305	134.5548
10	0.90	15.29	200	12	0.0600 ± 0.0168	8	0.6667	7	0.5833	0.0945	879549.6875
11	0.50	190.16	200	127	0.6350 ± 0.0340	80	0.6299	99	0.7795	0.0245	136.5050
12	0.14	664.90	200	21	0.1050 ± 0.0217	9	0.4286	3	0.1429	0.0191	1062.2715
13	0.50	163.57	200	105	0.5250 ± 0.0353	52	0.4952	71	0.6762	0.0264	136.4800
14	0.38	531.92	200	181	0.9050 ± 0.0207	114	0.6298	125	0.6906	0.0150	232.8503
15	0.14	1196.83	200	117	0.5850 ± 0.0348	67	0.5726	78	0.6667	0.0161	590.5674
16	0.50	110.37	200	46	0.2300 ± 0.0298	21	0.4565	22	0.4783	0.0300	135.2062
17	0.10	1276.62	200	74	0.3700 ± 0.0341	50	0.6757	32	0.4324	0.0152	751.2043
18	0.90	20.61	200	33	0.1650 ± 0.0262	26	0.7879	18	0.5455	0.0804	863.1549
19	0.66	132.98	200	168	0.8400 ± 0.0259	103	0.6131	129	0.7679	0.0316	123.8590
20	0.90	12.63	200	18	0.0900 ± 0.0202	11	0.6111	6	0.3333	0.0712	359.1638
21	0.90	17.95	200	26	0.1300 ± 0.0238	20	0.7692	11	0.4231	0.0876	170.1461
22	0.90	23.27	200	33	0.1650 ± 0.0262	26	0.7879	16	0.4848	0.0860	302.4461
23	0.78	79.79	200	147	0.7350 ± 0.0312	91	0.6190	103	0.7007	0.0494	159.9960
24	0.10	1542.58	200	111	0.5550 ± 0.0351	73	0.6577	69	0.6216	0.0145	770.1436

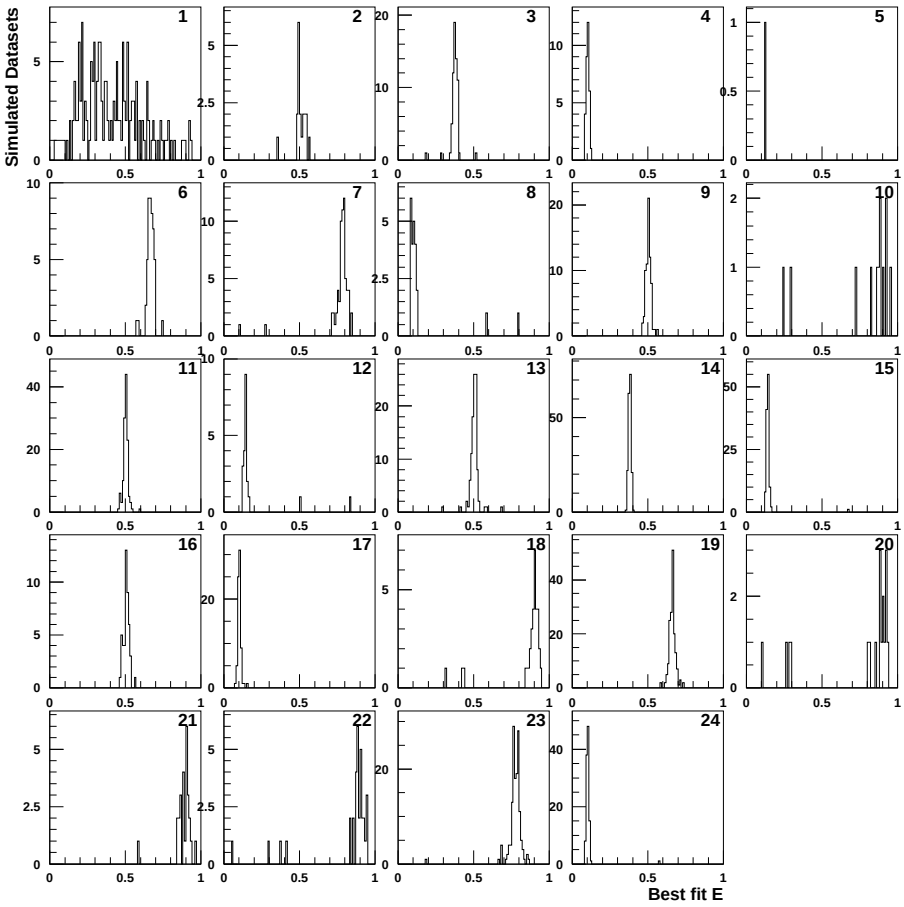

Figure 21: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Georgios claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

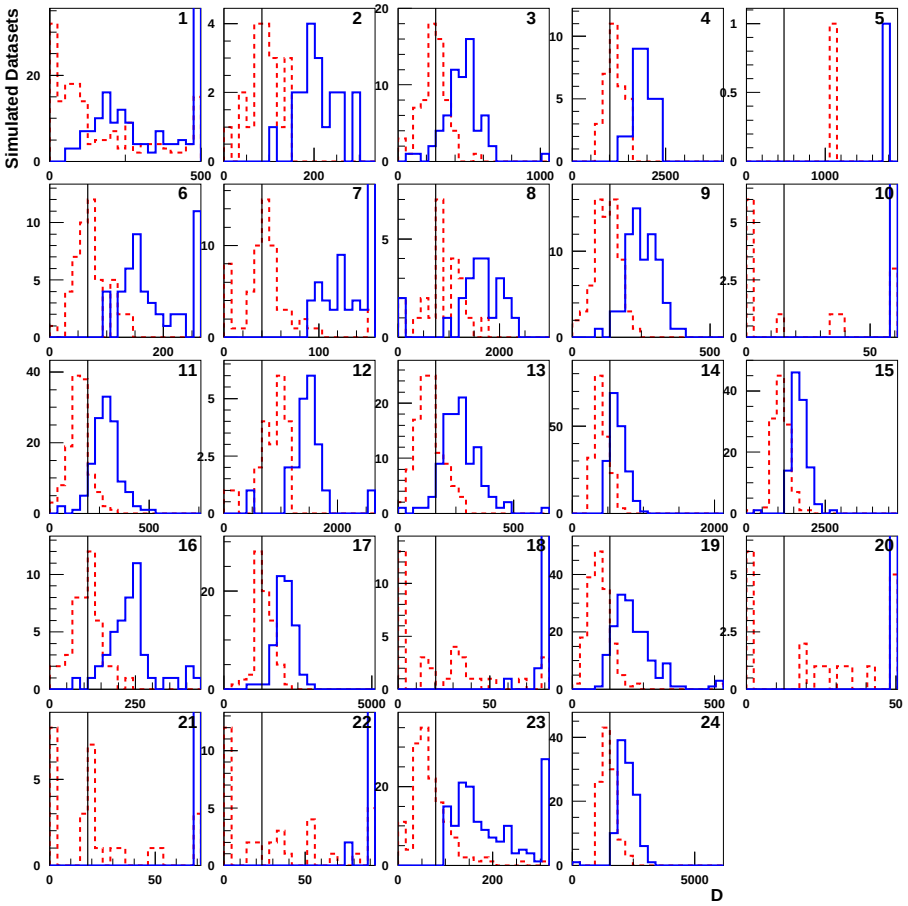

Figure 22: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Georgios claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

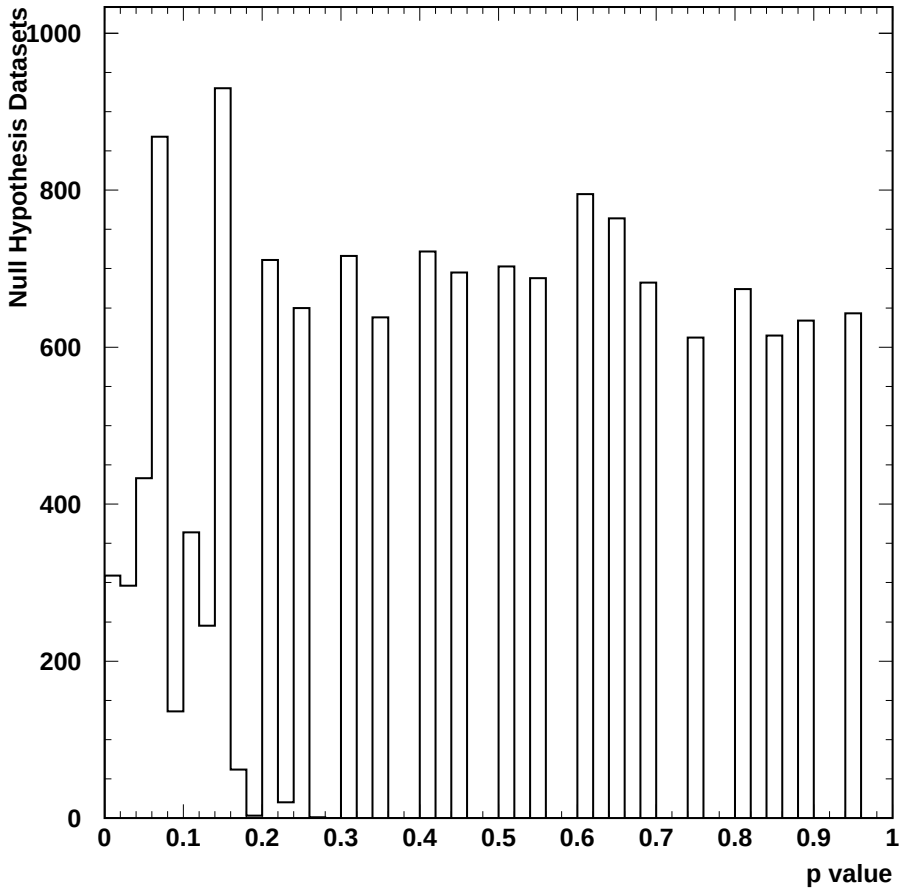

Figure 23: Distribution of the quoted p value in null hypothesis challenge datasets for Georgios's solution to Problem 1.

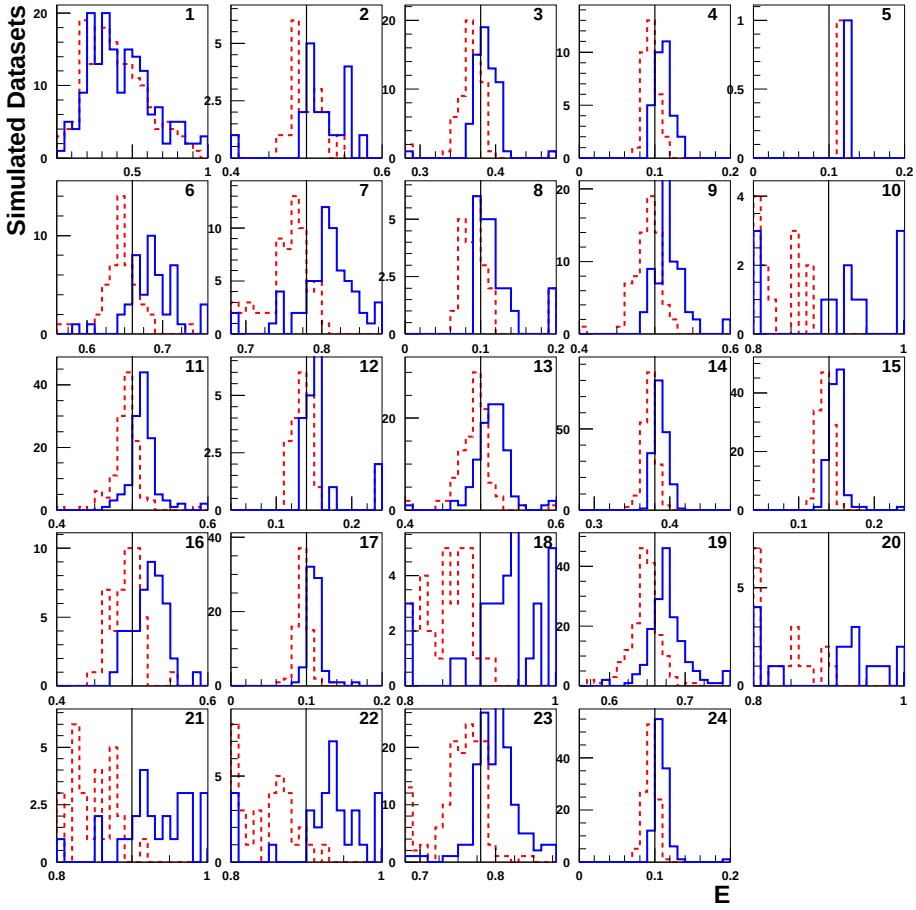

Figure 24: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Georgios claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.7 From Mark Allen

For Challenge Problem #1, Mark Allen provided a solution based on an unbinned maximum-likelihood fit, $\Delta \log \mathcal{L}$ as the test statistic for computing p values. In order to find the global maximum of the likelihood most often, several fits are performed with different starting conditions. The p values are computed by comparing a dataset's test statistic with a distribution of a large number of simulated background-only datasets. Since a signal can be found anywhere in the distribution on any of the simulated background-only datasets, the Look-Elsewhere Effect is taken into account.

Table 9 lists the error rates in the challenge datasets for Mark's solution. The Type-I error rate is not measurably different from 1%, as desired.

Table 9: Problem 1 performance evaluation for Mark Allen's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	$E_{\rm true}$	$D_{\rm true}$	$n_{\rm rep}$	$n_{\rm disc}$	$f_{\rm disc}$	$n_{\rm Ecorr}$	$f_{\rm Ecorr}$	$n_{\rm Dcorr}$	$f_{\rm Dcorr}$	⟨Ewid⟩	⟨Dwid⟩
1	—	0.00	15400	164	0.0106 ± 0.0008	—	—	0	0.0000	0.0256	377.2259
2	0.50	83.78	200	24	0.1200 ± 0.0230	16	0.6667	2	0.0833	0.0292	119.2689
3	0.38	265.96	200	114	0.5700 ± 0.0350	68	0.5965	74	0.6491	0.0228	196.1854
4	0.10	1010.65	200	80	0.4000 ± 0.0346	51	0.6375	44	0.5500	0.0187	736.7384
5	0.10	478.73	200	5	0.0250 ± 0.0110	2	0.4000	0	0.0000	0.0197	752.8884
6	0.66	66.49	200	66	0.3300 ± 0.0332	42	0.6364	43	0.6515	0.0286	80.7764
7	0.78	39.89	200	75	0.3750 ± 0.0342	51	0.6800	41	0.5467	0.0310	72.8878
8	0.10	744.69	200	40	0.2000 ± 0.0283	18	0.4500	5	0.1250	0.0192	704.9865
9	0.50	136.97	200	105	0.5250 ± 0.0353	65	0.6190	62	0.5905	0.0244	129.4742
10	0.90	15.29	200	16	0.0800 ± 0.0192	7	0.4375	0	0.0000	0.0369	49.9948
11	0.50	190.16	200	160	0.8000 ± 0.0283	111	0.6938	130	0.8125	0.0224	134.1713
12	0.14	664.90	200	50	0.2500 ± 0.0306	28	0.5600	13	0.2600	0.0205	562.1276
13	0.50	163.57	200	130	0.6500 ± 0.0337	80	0.6154	87	0.6692	0.0233	130.9396
14	0.38	531.92	200	199	0.9950 ± 0.0050	136	0.6834	149	0.7487	0.0153	222.7198
15	0.14	1196.83	200	185	0.9250 ± 0.0186	127	0.6865	146	0.7892	0.0169	586.5714
16	0.50	110.37	200	69	0.3450 ± 0.0336	38	0.5507	29	0.4203	0.0268	126.2137
17	0.10	1276.62	200	136	0.6800 ± 0.0330	88	0.6471	95	0.6985	0.0171	744.3584
18	0.90	20.61	200	54	0.2700 ± 0.0314	37	0.6852	31	0.5741	0.0355	63.1787
19	0.66	132.98	200	184	0.9200 ± 0.0192	111	0.6033	147	0.7989	0.0232	95.3406
20	0.90	12.63	200	27	0.1350 ± 0.0242	18	0.6667	0	0.0000	0.0355	78.6395
21	0.90	17.95	200	42	0.2100 ± 0.0288	25	0.5952	1	0.0238	0.0352	57.5532
22	0.90	23.27	200	53	0.2650 ± 0.0312	37	0.6981	24	0.4528	0.0353	53.9548
23	0.78	79.79	200	164	0.8200 ± 0.0272	110	0.6707	134	0.8171	0.0273	74.6528
24	0.10	1542.58	200	178	0.8900 ± 0.0221	125	0.7022	135	0.7584	0.0155	773.4517

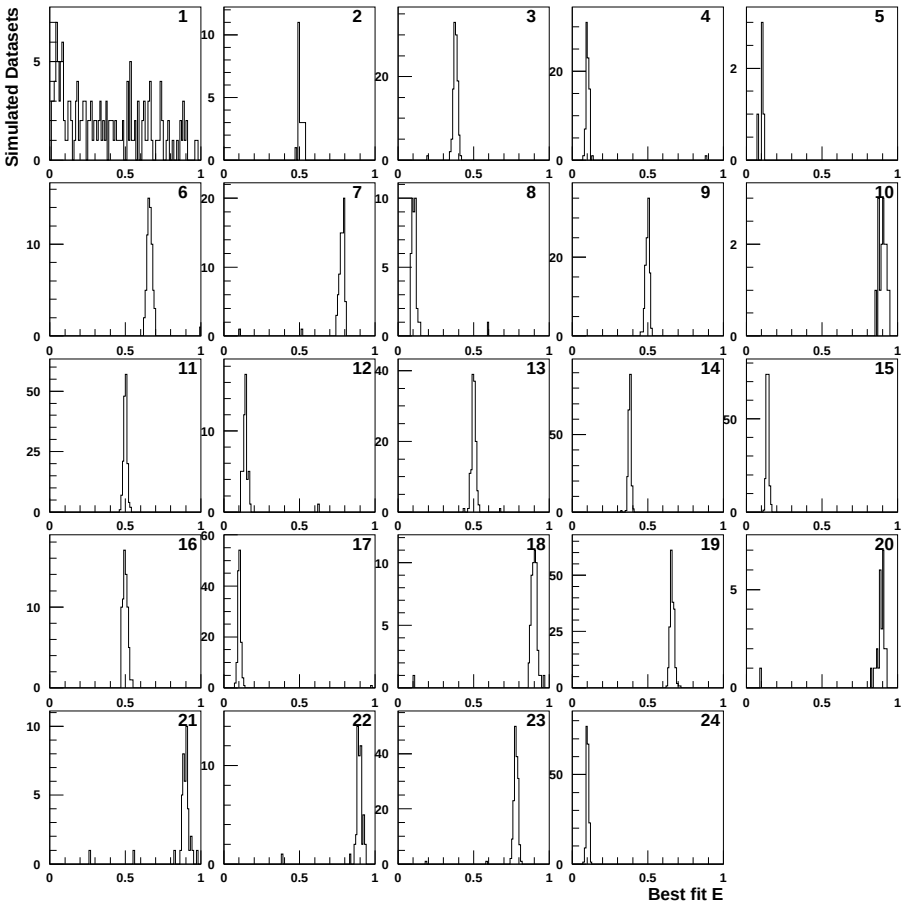

Figure 25: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Mark claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

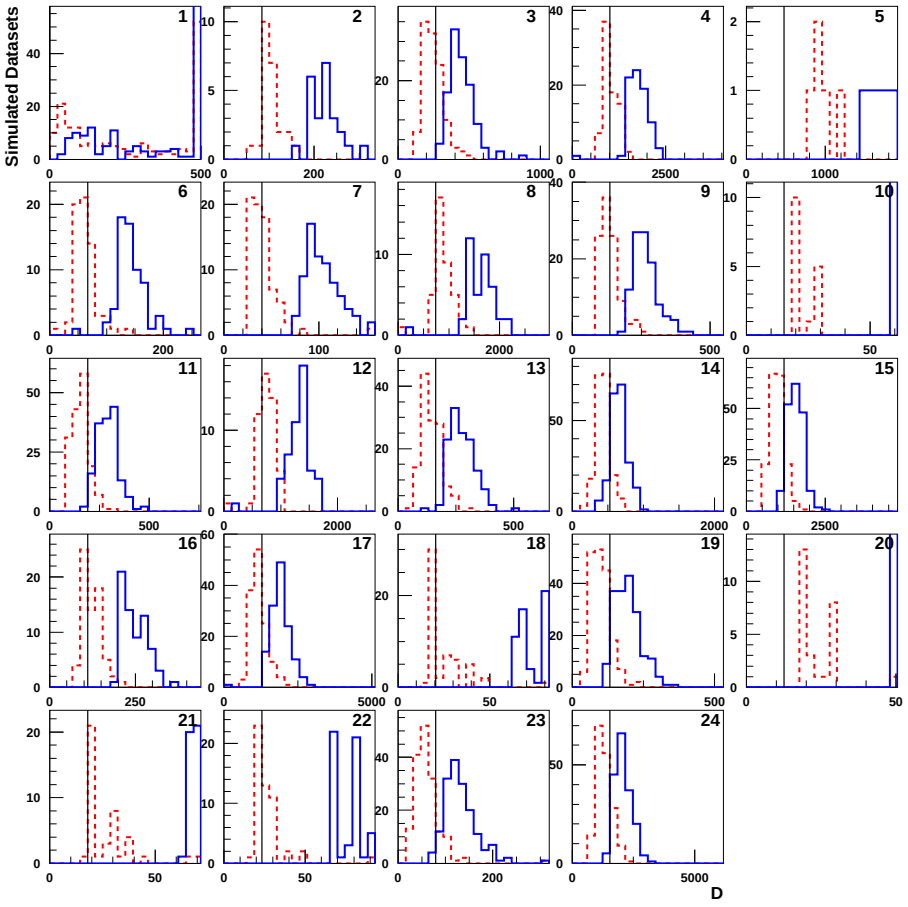

Figure 26: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Mark claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

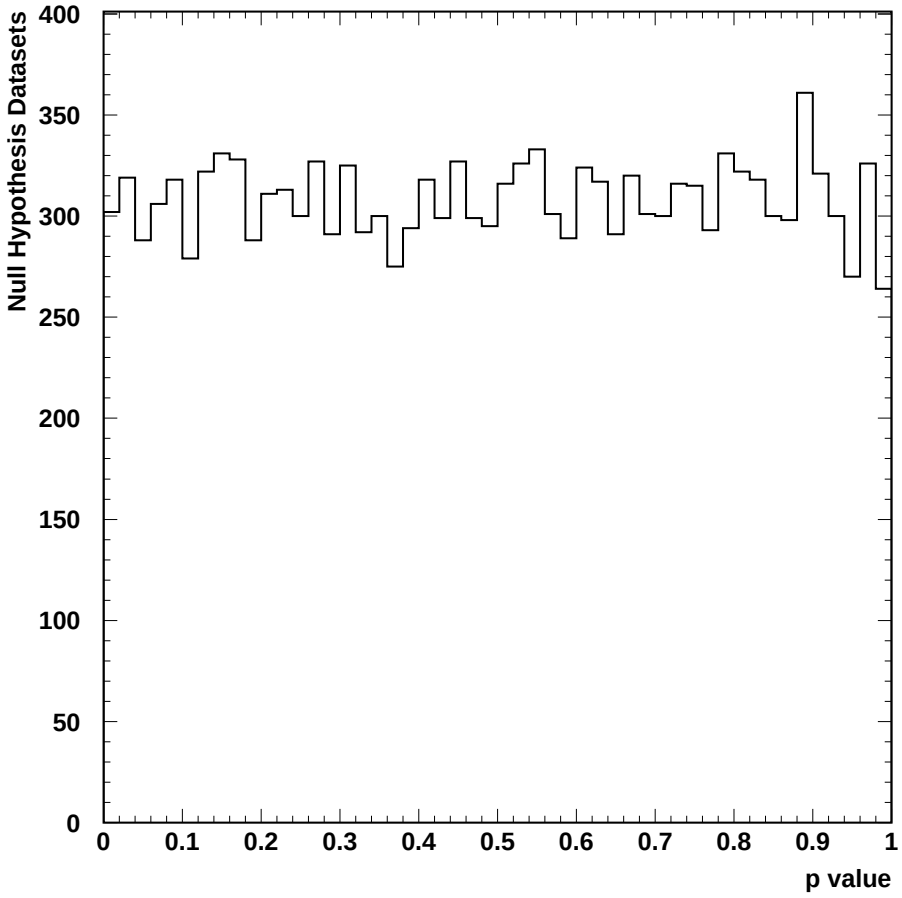

Figure 27: Distribution of the quoted p value in null hypothesis challenge datasets for Mark's solution to Problem 1.

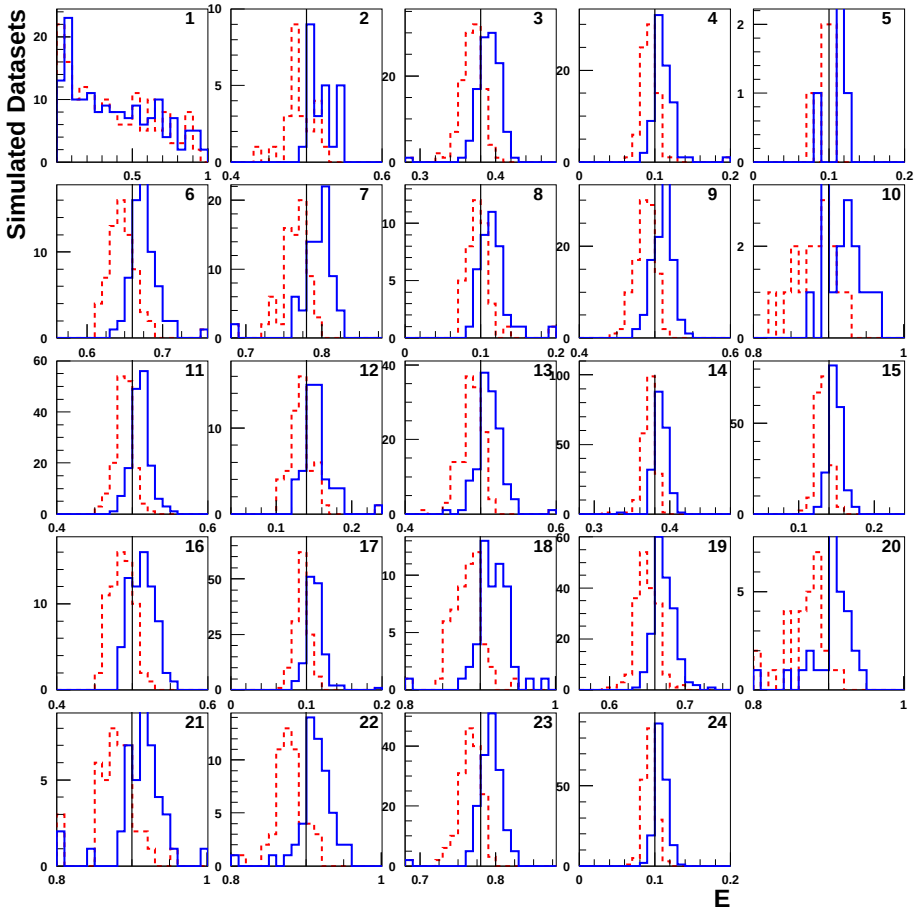

Figure 28: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Mark claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.8 From Frederik Beaujean and the BAT Team

The Bayesian Analysis Toolkit (BAT) Team consists of F. Beaujean, A. Caldwell, and S. Pashapour. For Challenge Problem #1, Frederik Beaujean provided a solution based on the BAT's fast Poisson p value estimation, corrected for the number of degrees of freedom. The value of A that maximizes the posterior probability in the background-only case is used. If the p-value is less than 0.01, a Bayesian analysis is conducted, and a discovery is claimed if $P(B|\text{Data}) < 0.001$. The Look-Elsewhere Effect is taken into account by assuming a prior that favors the background model. A rather small fraction of the simulated datasets with injected signals had a discovery claim using this technique.

Table 10 lists the error rates in the challenge datasets for Frederik's solution. The Type-I error rate is not measurably different from zero given the size of the sample of simulated datasets. An earlier version of this note had misinterpreted the p value as the discovery choice, when in fact the additional requirement on the value of $P(B|\text{Data})$ lowers the false discovery rate to zero.

Table 10: Problem 1 performance evaluation for Frederik Beaujean's solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	⟨Ewid⟩	⟨Dwid⟩
1	—	0.00	15400	0	0.0000 ± 0.0000	—	—	0	—	—	—
2	0.5	83.78	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
3	0.4	265.96	200	7	0.0350 ± 0.0130	3	0.4286	0	0.0000	0.0178	219.7268
4	0.1	1010.65	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
5	0.1	478.73	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
6	0.7	66.49	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
7	0.8	39.89	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
8	0.1	744.69	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
9	0.5	136.97	200	6	0.0300 ± 0.0121	4	0.6667	0	0.0000	0.0174	150.6918
10	0.9	15.29	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
11	0.5	190.16	200	25	0.1250 ± 0.0234	17	0.6800	4	0.1600	0.0184	148.2737
12	0.1	664.90	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
13	0.5	163.57	200	15	0.0750 ± 0.0186	8	0.5333	0	0.0000	0.0190	148.2983
14	0.4	531.92	200	124	0.6200 ± 0.0343	86	0.6935	97	0.7823	0.0146	227.3591
15	0.1	1196.83	200	4	0.0200 ± 0.0099	1	0.2500	1	0.2500	0.0123	454.7875
16	0.5	110.37	200	1	0.0050 ± 0.0050	1	1.0000	0	0.0000	0.0201	148.3780
17	0.1	1276.62	200	1	0.0050 ± 0.0050	1	1.0000	0	0.0000	0.0119	457.0601
18	0.9	20.61	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
19	0.7	132.98	200	50	0.2500 ± 0.0306	27	0.5400	27	0.5400	0.0198	106.3467
20	0.9	12.63	200	0	0.0000 ± 0.0000	0	—	0	—	—	—
21	0.9	17.95	200	1	0.0050 ± 0.0050	0	0.0000	0	0.0000	0.0322	62.6962
22	0.9	23.27	200	3	0.0150 ± 0.0086	1	0.3333	0	0.0000	0.0260	67.1801
23	0.8	79.79	200	33	0.1650 ± 0.0262	22	0.6667	7	0.2121	0.0216	87.0117
24	0.1	1542.58	200	1	0.0050 ± 0.0050	0	0.0000	1	1.0000	0.0134	534.8101

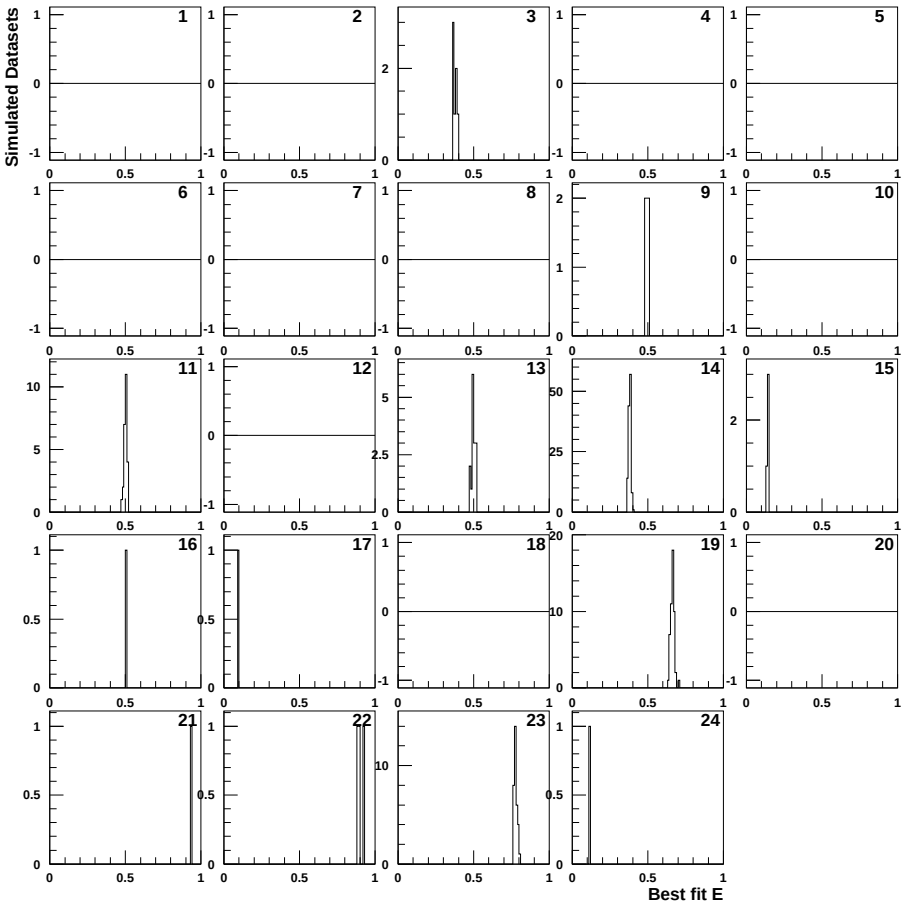

Figure 29: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Frederik claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

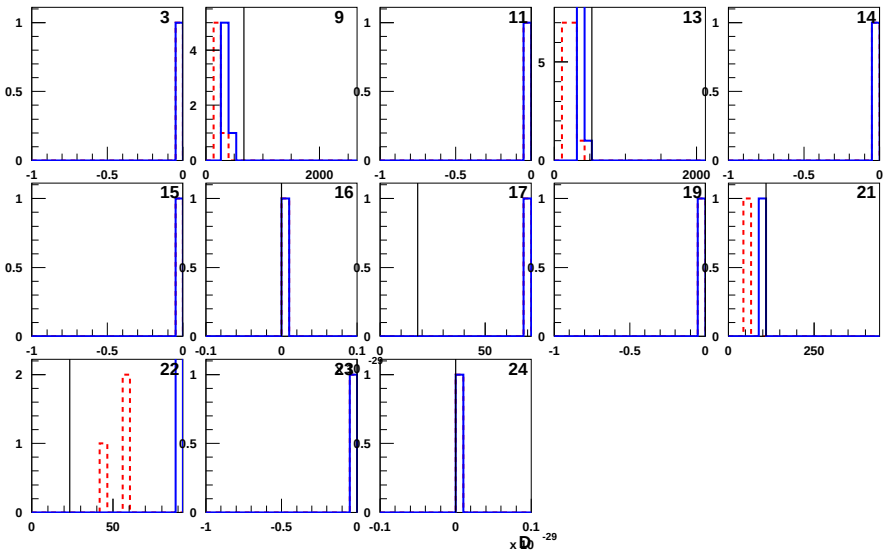

Figure 30: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Frederik claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels. Only those signal categories with at least one simulated dataset resulting in a discovery are shown.

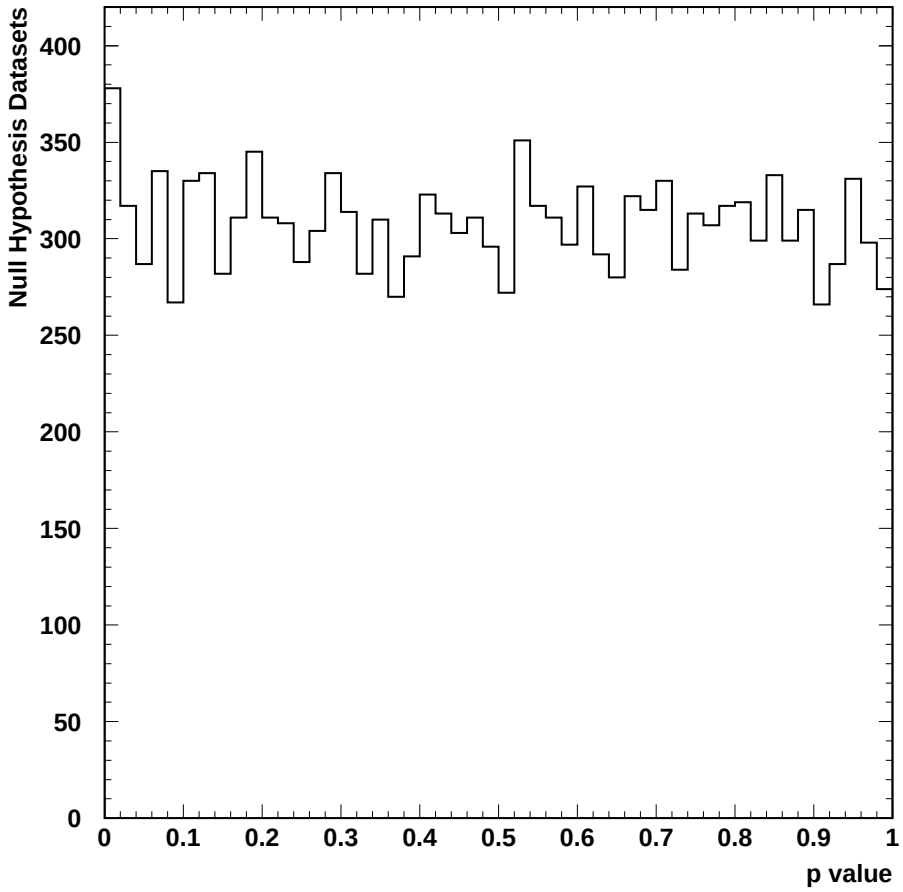

Figure 31: Distribution of the quoted p value in null hypothesis challenge datasets for Frederik's solution to Problem 1. Outcomes with $p < 0.01$ are subject to a further requirement that $p(B|\text{Data}) < 0.001$.

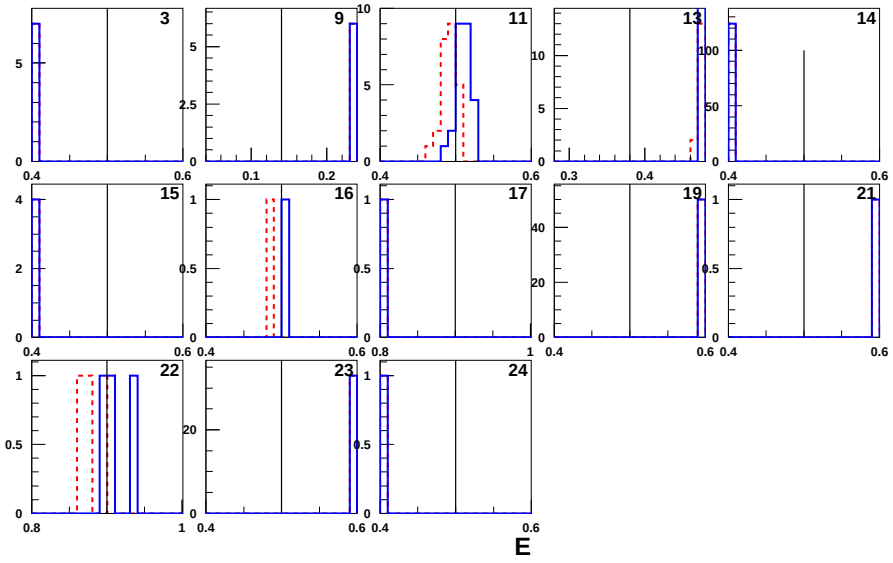

Figure 32: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Frederik claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Only those signal categories with at least one simulated dataset resulting in a discovery are shown.

2.2.9 From Stefan Schmitt

For Challenge Problem #1, Stefan Schmitt provided a solution based on a fractional event-counting procedure. Two solutions were provided, one using unbinned fits to the data, while the other bins the data. The weights in Stefan's method are designed to test for a signal at a particular value of E and to suppress contributions far from the tested peak. Stefan thus scans E in fine steps to find the best value of E – the one with the lowest p value. The Look-Elsewhere Effect thus pushes the mean of the p value distribution downwards, as can be seen in Figure 35 for his unbinned search. Stefan corrects these by running a Monte Carlo simulation of the null hypothesis and seeking a peak at all E in each one, making a distribution of the LEE-biased p values, and evaluates a new cut that gives an expected global error rate of 1%. He also corrects the p values using this Monte Carlo simulation – it is these corrected p values that are shown in Figures 36 and 40.

Table 11 lists the error rates in the challenge datasets for Stefan's unbinned solution, and Table 12 lists the same information for Stefan's binned solution. The Type-I error rates are not measurably different from 1%, as desired. The performances are similar between the binned submission and the unbinned submission.

Table 11: Problem 1 performance evaluation for Stefan Schmitt's unbinned solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	$\langle$Ewid$\rangle$	$\langle$Dwid$\rangle$
1	—	0.00	15400	172	0.0112 ± 0.0008	—	—	40	0.2326	0.0419	502.3421
2	0.50	83.78	200	24	0.1200 ± 0.0230	17	0.7083	21	0.8750	0.0481	91.8503
3	0.38	265.96	200	114	0.5700 ± 0.0350	98	0.8596	65	0.5702	0.0418	181.2448
4	0.10	1010.65	200	90	0.4500 ± 0.0352	79	0.8778	50	0.5556	0.0358	829.2986
5	0.10	478.73	200	9	0.0450 ± 0.0147	6	0.6667	7	0.7778	0.0383	902.7023
6	0.66	66.49	200	69	0.3450 ± 0.0336	57	0.8261	43	0.6232	0.0428	48.0288
7	0.78	39.89	200	73	0.3650 ± 0.0340	50	0.6849	50	0.6849	0.0464	44.0940
8	0.10	744.69	200	30	0.1500 ± 0.0252	25	0.8333	23	0.7667	0.0367	829.5305
9	0.50	136.97	200	109	0.5450 ± 0.0352	96	0.8807	70	0.6422	0.0422	101.0201
10	0.90	15.29	200	19	0.0950 ± 0.0207	6	0.3158	13	0.6842	0.0410	22.8564
11	0.50	190.16	200	156	0.7800 ± 0.0293	141	0.9038	87	0.5577	0.0391	104.3673
12	0.14	664.90	200	45	0.2250 ± 0.0295	34	0.7556	37	0.8222	0.0384	653.5959
13	0.50	163.57	200	133	0.6650 ± 0.0334	111	0.8346	62	0.4662	0.0405	100.0957
14	0.38	531.92	200	199	0.9950 ± 0.0050	188	0.9447	97	0.4874	0.0318	184.5835
15	0.14	1196.83	200	182	0.9100 ± 0.0202	171	0.9396	84	0.4615	0.0347	638.5347
16	0.50	110.37	200	65	0.3250 ± 0.0331	50	0.7692	51	0.7846	0.0443	98.8377
17	0.10	1276.62	200	140	0.7000 ± 0.0324	118	0.8429	72	0.5143	0.0348	814.2061
18	0.90	20.61	200	55	0.2750 ± 0.0316	30	0.5455	35	0.6364	0.0455	22.4458
19	0.66	132.98	200	184	0.9200 ± 0.0192	142	0.7717	76	0.4130	0.0375	60.1174
20	0.90	12.63	200	24	0.1200 ± 0.0230	13	0.5417	13	0.5417	0.0486	55.7148
21	0.90	17.95	200	37	0.1850 ± 0.0275	17	0.4595	21	0.5676	0.0440	24.9794
22	0.90	23.27	200	50	0.2500 ± 0.0306	30	0.6000	44	0.8800	0.0443	23.1247
23	0.78	79.79	200	157	0.7850 ± 0.0290	119	0.7580	67	0.4268	0.0406	41.1542
24	0.10	1542.58	200	178	0.8900 ± 0.0221	168	0.9438	82	0.4607	0.0319	832.7157

Table 12: Problem 1 performance evaluation for Stefan Schmitt's binned solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	$\langle$Ewid$\rangle$	$\langle$Dwid$\rangle$
1	—	0.00	15400	169	0.0110 ± 0.0008	—	—	0	0.0000	0.0416	397.1595
2	0.50	83.78	200	21	0.1050 ± 0.0217	16	0.7619	14	0.6667	0.0477	119.0175
3	0.38	265.96	200	123	0.6150 ± 0.0344	105	0.8537	80	0.6504	0.0426	192.0853
4	0.10	1010.65	200	77	0.3850 ± 0.0344	69	0.8961	55	0.7143	0.0355	766.9265
5	0.10	478.73	200	7	0.0350 ± 0.0130	5	0.7143	2	0.2857	0.0400	714.8434
6	0.66	66.49	200	71	0.3550 ± 0.0338	60	0.8451	55	0.7746	0.0453	75.2635
7	0.78	39.89	200	70	0.3500 ± 0.0337	50	0.7143	47	0.6714	0.0465	68.3524
8	0.10	744.69	200	34	0.1700 ± 0.0266	27	0.7941	20	0.5882	0.0363	761.5689
9	0.50	136.97	200	109	0.5450 ± 0.0352	98	0.8991	82	0.7523	0.0425	127.4503
10	0.90	15.29	200	24	0.1200 ± 0.0230	10	0.4167	17	0.7083	0.0429	43.2923
11	0.50	190.16	200	157	0.7850 ± 0.0290	144	0.9172	118	0.7516	0.0397	134.7139
12	0.14	664.90	200	47	0.2350 ± 0.0300	36	0.7660	30	0.6383	0.0384	595.2405
13	0.50	163.57	200	129	0.6450 ± 0.0338	107	0.8295	86	0.6667	0.0406	129.2215
14	0.38	531.92	200	199	0.9950 ± 0.0050	189	0.9497	144	0.7236	0.0321	222.3337
15	0.14	1196.83	200	180	0.9000 ± 0.0212	171	0.9500	118	0.6556	0.0351	588.5422
16	0.50	110.37	200	68	0.3400 ± 0.0335	57	0.8382	46	0.6765	0.0456	125.5051
17	0.10	1276.62	200	135	0.6750 ± 0.0331	116	0.8593	91	0.6741	0.0344	734.1825
18	0.90	20.61	200	59	0.2950 ± 0.0322	29	0.4915	37	0.6271	0.0461	48.6783
19	0.66	132.98	200	185	0.9250 ± 0.0186	145	0.7838	124	0.6703	0.0380	95.6798
20	0.90	12.63	200	26	0.1300 ± 0.0238	13	0.5000	9	0.3462	0.0497	71.5992
21	0.90	17.95	200	44	0.2200 ± 0.0293	23	0.5227	27	0.6136	0.0449	47.2139
22	0.90	23.27	200	59	0.2950 ± 0.0322	33	0.5593	38	0.6441	0.0453	48.3993
23	0.78	79.79	200	156	0.7800 ± 0.0293	125	0.8013	111	0.7115	0.0410	73.0747
24	0.10	1542.58	200	178	0.8900 ± 0.0221	171	0.9607	118	0.6629	0.0319	757.0254

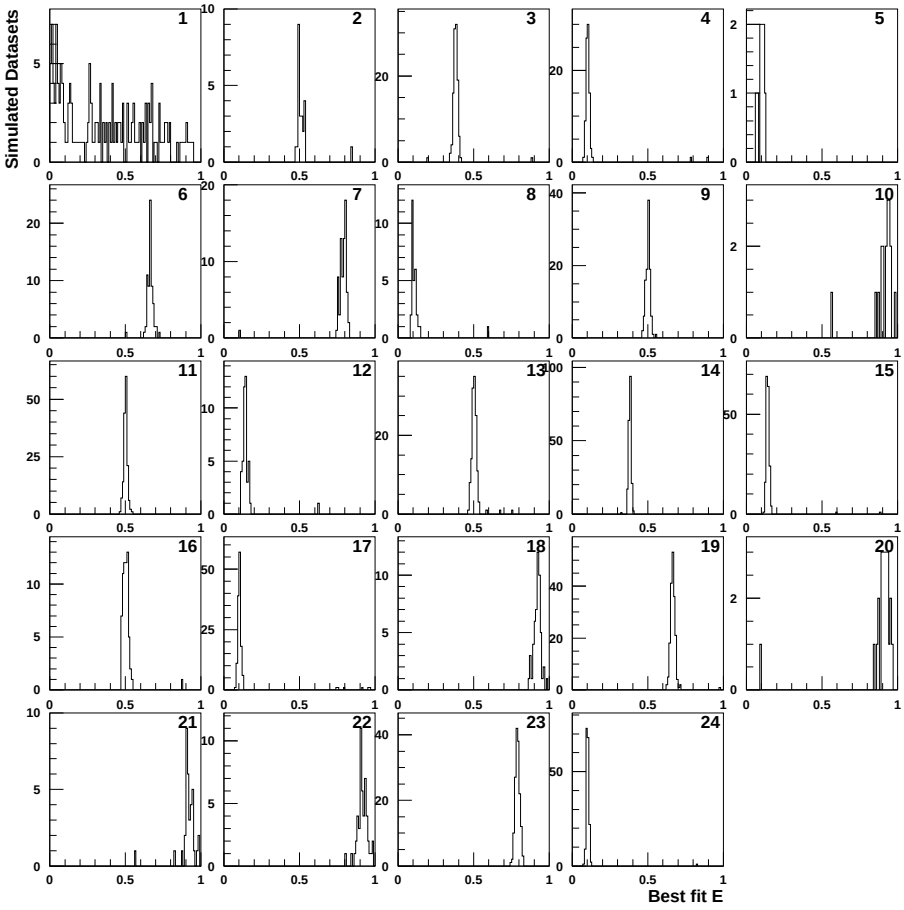

Figure 33: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his unbinned solution. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

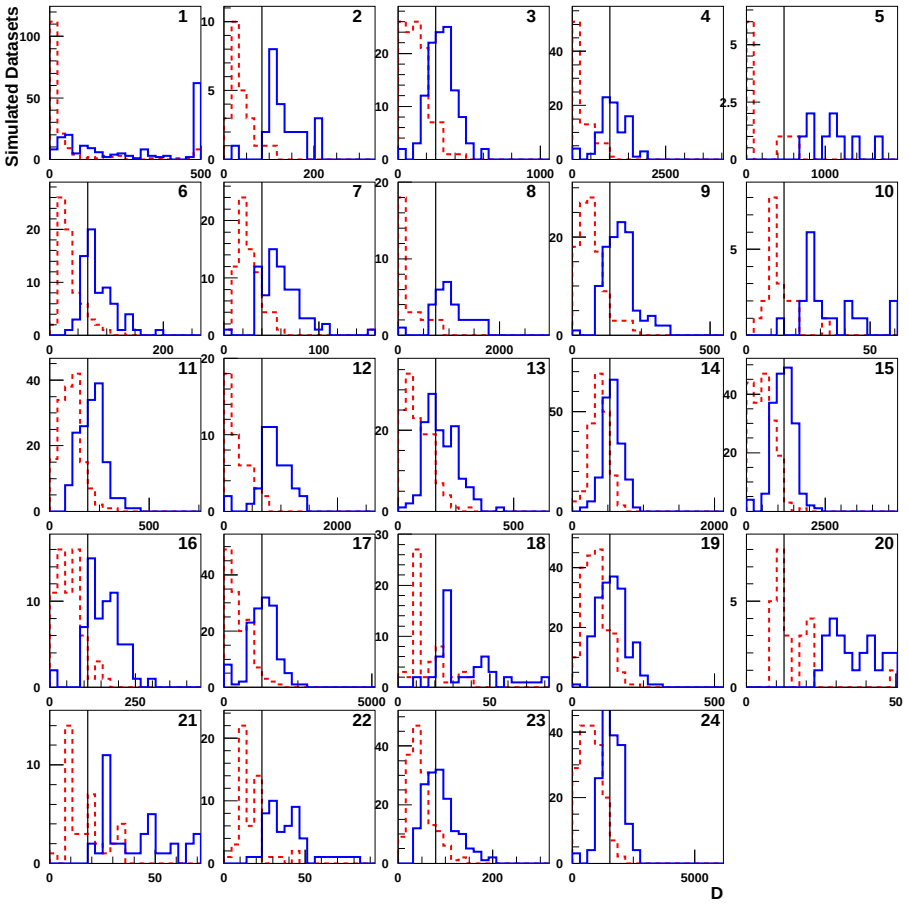

Figure 34: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his unbinned solution. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

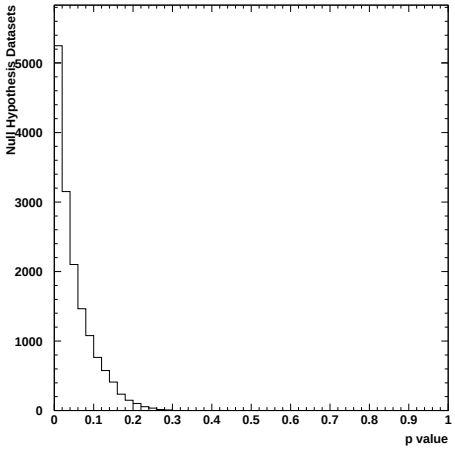

Figure 35: Distribution of the quoted p value in null hypothesis challenge datasets for Stefan's unbinned solution to Problem 1, before correcting for the Look-Elsewhere Effect.

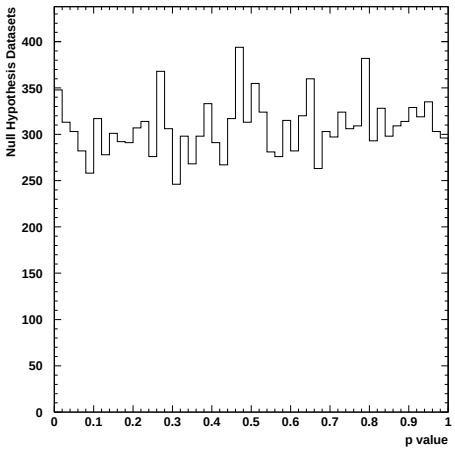

Figure 36: Distribution of the quoted p value in null hypothesis challenge datasets for Stefan's unbinned solution to Problem 1, after correcting for the Look-Elsewhere Effect using a Monte Carlo simulation.

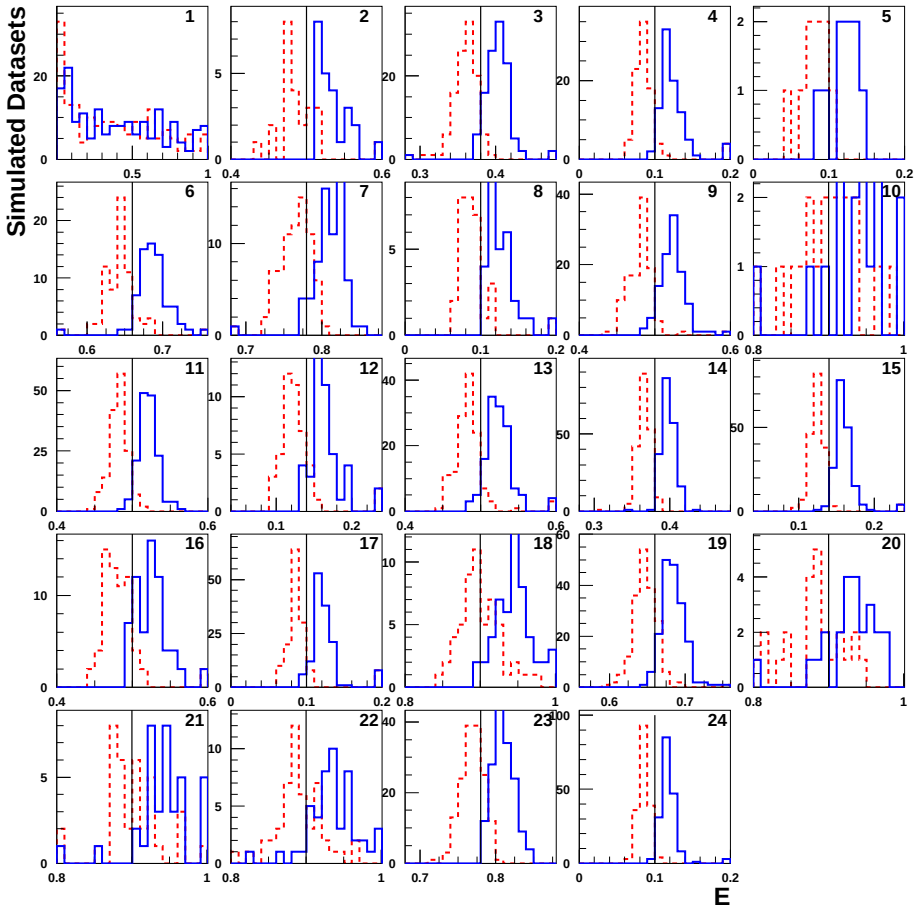

Figure 37: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his unbinned solution. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

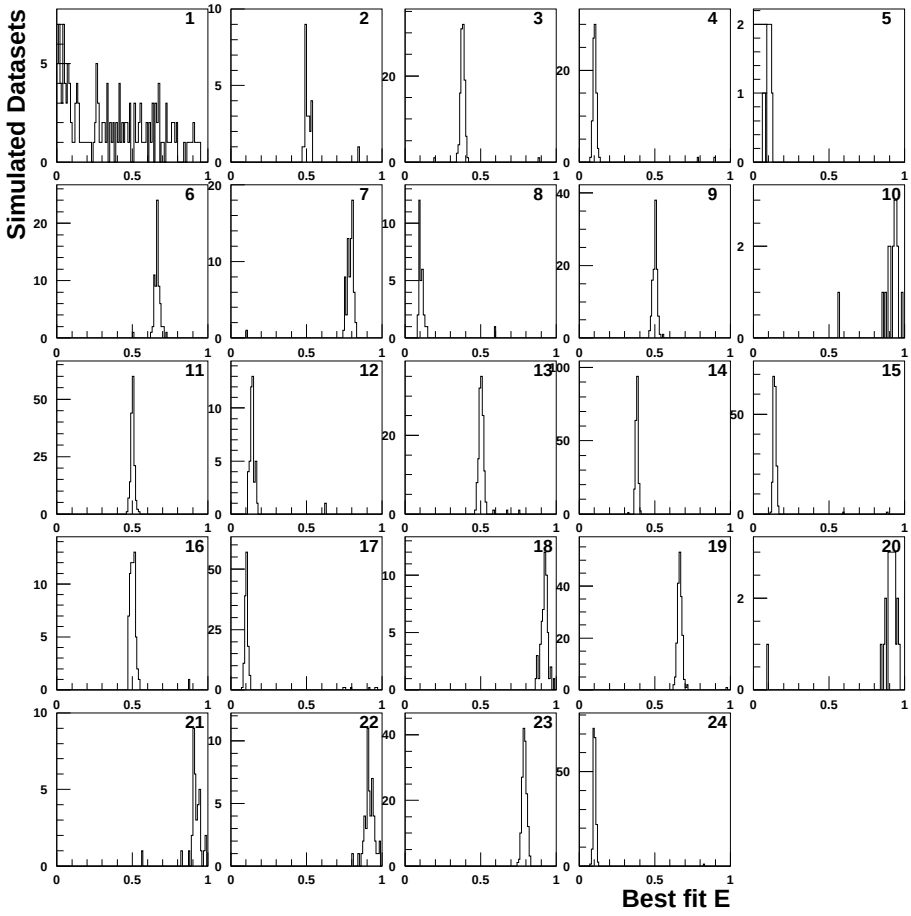

Figure 38: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his binned solution. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

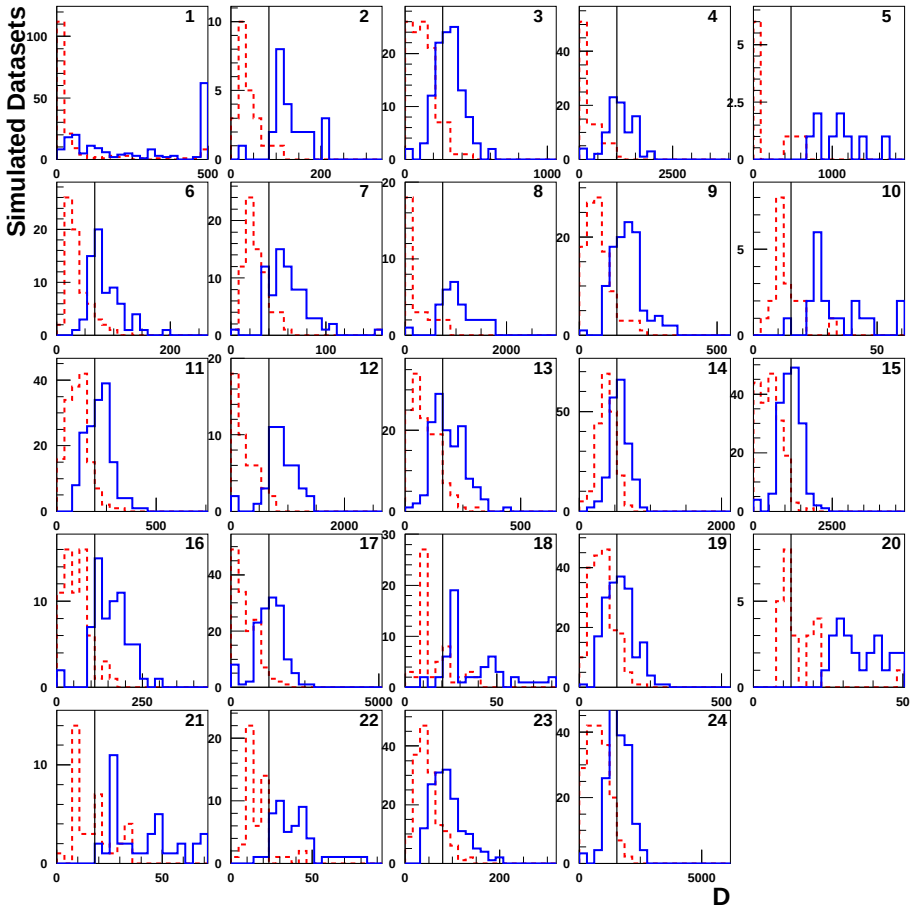

Figure 39: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his binned solution. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

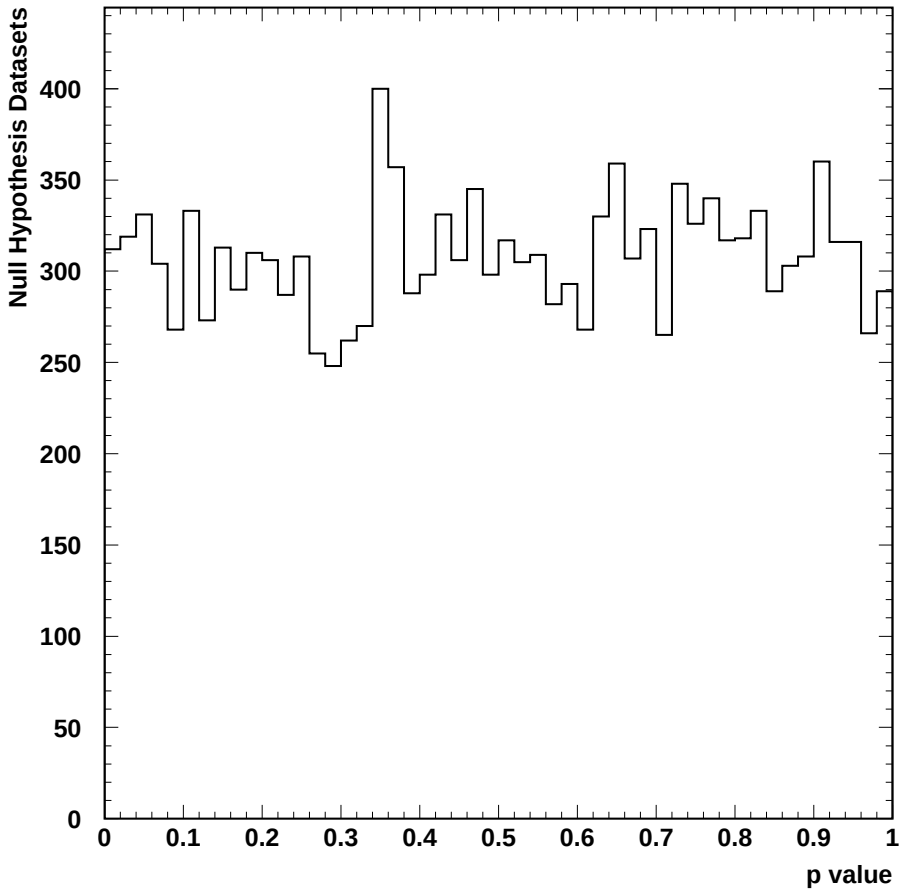

Figure 40: Distribution of the quoted p value in null hypothesis challenge datasets for Stefan's binned solution to Problem 1.

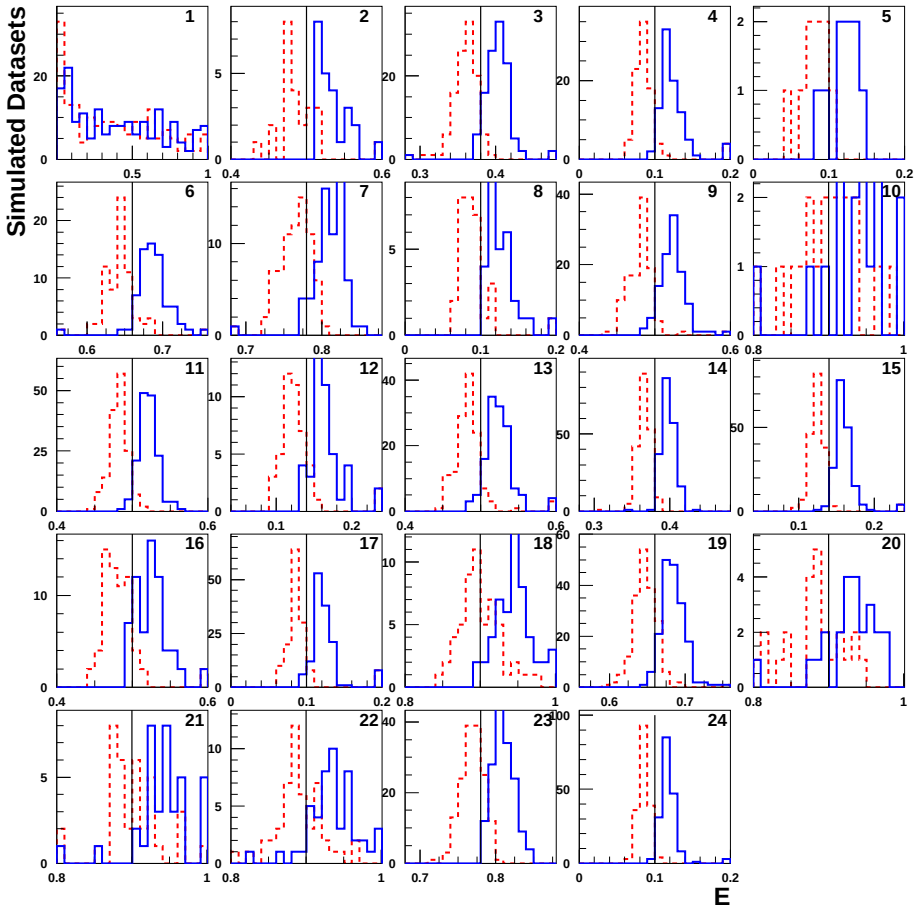

Figure 41: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his binned solution. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.2.10 From Stefano Andreon

For Challenge Problem #1, Stefano Andreon provided a solution based on a Bayesian computation with improper uniform priors on A and D, with a zero value for the prior for negative (unphysical) values, and an uniform prior, between 0 and 1, on E. Stefano computes $p(D = 0|\text{data})$, up to a multiplicative factor, and selects simulated datasets for discovery claims if $p(D = 0|\text{data}) < 3 \times 10^{-3}$ for the first solution, and $p(D = 0|\text{data}) < 4 \times 10^{-3}$ for the second. The Type-I error rate will be higher for the second set, but the power will also be larger. Stefano did not compute the power of his test. Due to time and computing limitations, Stefano analyzed only the first 10000 simulated datasets of Problem 1.

Table 13 lists the error rates in the challenge datasets for Stefano's $p(D = 0|\text{data}) < 3 \times 10^{-3}$ solution, and Table 14 lists the same information for Stefano's $p(D = 0|\text{data}) < 4 \times 10^{-3}$ solution. In both cases, the Type-I error rates exceed 1%, although for the solution with $p(D = 0|\text{data}) < 3 \times 10^{-3}$ solution, the significance of the claim that the Type-I error rate is too high is only $\sim 2\sigma$.

Table 13: Problem 1 performance evaluation for Stefano Andreon's $p(D = 0|\text{data}) < 3 \times 10^{-3}$ solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	⟨Ewid⟩	⟨Dwid⟩
1	—	0.00	7685	97	0.0126 ± 0.0013	—	—	0	0.0000	0.0361	413.9379
2	0.50	83.78	94	8	0.0851 ± 0.0288	5	0.6250	0	0.0000	0.0365	152.4175
3	0.38	265.96	103	65	0.6311 ± 0.0475	44	0.6769	50	0.7692	0.0266	207.3465
4	0.10	1010.65	106	51	0.4811 ± 0.0485	41	0.8039	48	0.9412	0.0260	693.8154
5	0.10	478.73	103	6	0.0583 ± 0.0231	3	0.5000	2	0.3333	0.0307	684.7844
6	0.66	66.49	101	19	0.1881 ± 0.0389	14	0.7368	11	0.5789	0.0324	89.5388
7	0.78	39.89	96	13	0.1354 ± 0.0349	7	0.5385	0	0.0000	0.0310	122.8128
8	0.10	744.69	103	27	0.2621 ± 0.0433	18	0.6667	25	0.9259	0.0394	655.4998
9	0.50	136.97	107	51	0.4766 ± 0.0483	29	0.5686	35	0.6863	0.0353	136.4956
10	0.90	15.29	110	0	0.0000 ± 0.0000	0	—	0	—	—	—
11	0.50	190.16	105	86	0.8190 ± 0.0376	59	0.6860	77	0.8953	0.0246	138.1537
12	0.14	664.90	105	42	0.4000 ± 0.0478	32	0.7619	38	0.9048	0.0282	561.0679
13	0.50	163.57	112	69	0.6161 ± 0.0460	42	0.6087	53	0.7681	0.0343	137.7171
14	0.38	531.92	96	96	1.0000 ± 0.0000	66	0.6875	63	0.6562	0.0165	229.3821
15	0.14	1196.83	111	108	0.9730 ± 0.0154	82	0.7593	60	0.5556	0.0215	575.5975
16	0.50	110.37	101	28	0.2772 ± 0.0445	17	0.6071	14	0.5000	0.0295	129.5641
17	0.10	1276.62	89	68	0.7640 ± 0.0450	46	0.6765	42	0.6176	0.0237	700.2609
18	0.90	20.61	88	0	0.0000 ± 0.0000	0	—	0	—	—	—
19	0.66	132.98	106	95	0.8962 ± 0.0296	62	0.6526	81	0.8526	0.0250	97.2977
20	0.90	12.63	101	1	0.0099 ± 0.0099	0	0.0000	0	0.0000	0.0225	737.3783
21	0.90	17.95	83	1	0.0120 ± 0.0120	1	1.0000	0	0.0000	0.0317	69.6021
22	0.90	23.27	107	2	0.0187 ± 0.0131	1	0.5000	0	0.0000	0.0328	72.0091
23	0.78	79.79	98	65	0.6633 ± 0.0477	49	0.7538	52	0.8000	0.0288	80.2345
24	0.10	1542.58	91	82	0.9011 ± 0.0313	66	0.8049	29	0.3537	0.0202	705.2245

Table 14: Problem 1 performance evaluation for Stefano Andreon's $p(D = 0|\text{data}) < 4 \times 10^{-3}$ solution. The uncertainties quoted are the Gaussian approximation to binomial uncertainties, $\sqrt{f(1-f)/n}$. See the caption of Table 3 for definitions of the columns.

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	⟨Ewid⟩	⟨Dwid⟩
1	—	0.00	7685	147	0.0191 ± 0.0016	—	—	0	0.0000	0.0426	438.8728
2	0.50	83.78	94	11	0.1170 ± 0.0332	8	0.7273	2	0.1818	0.0664	149.4704
3	0.38	265.96	103	69	0.6699 ± 0.0463	48	0.6957	54	0.7826	0.0289	207.4382
4	0.10	1010.65	106	55	0.5189 ± 0.0485	44	0.8000	52	0.9455	0.0266	699.9116
5	0.10	478.73	103	7	0.0680 ± 0.0248	3	0.4286	3	0.4286	0.0315	664.8088
6	0.66	66.49	101	23	0.2277 ± 0.0417	18	0.7826	15	0.6522	0.0472	91.3774
7	0.78	39.89	96	18	0.1875 ± 0.0398	11	0.6111	0	0.0000	0.0330	112.8416
8	0.10	744.69	103	37	0.3592 ± 0.0473	26	0.7027	34	0.9189	0.0502	668.2768
9	0.50	136.97	107	51	0.4766 ± 0.0483	29	0.5686	35	0.6863	0.0353	136.4956
10	0.90	15.29	110	0	0.0000 ± 0.0000	0	—	0	—	—	—
11	0.50	190.16	105	86	0.8190 ± 0.0376	59	0.6860	77	0.8953	0.0246	138.1537
12	0.14	664.90	105	46	0.4381 ± 0.0484	36	0.7826	42	0.9130	0.0298	568.0446
13	0.50	163.57	112	72	0.6429 ± 0.0453	43	0.5972	56	0.7778	0.0343	137.2235
14	0.38	531.92	96	96	1.0000 ± 0.0000	66	0.6875	63	0.6562	0.0165	229.3821
15	0.14	1196.83	111	109	0.9820 ± 0.0126	83	0.7615	60	0.5505	0.0216	575.6018
16	0.50	110.37	101	29	0.2871 ± 0.0450	17	0.5862	15	0.5172	0.0441	144.0979
17	0.10	1276.62	89	71	0.7978 ± 0.0426	49	0.6901	42	0.5915	0.0241	702.0446
18	0.90	20.61	88	2	0.0227 ± 0.0159	1	0.5000	0	0.0000	0.0394	372.0735
19	0.66	132.98	106	97	0.9151 ± 0.0271	63	0.6495	83	0.8557	0.0253	97.3551
20	0.90	12.63	101	2	0.0198 ± 0.0139	0	0.0000	0	0.0000	0.0271	452.8327
21	0.90	17.95	83	1	0.0120 ± 0.0120	1	1.0000	0	0.0000	0.0317	69.6021
22	0.90	23.27	107	2	0.0187 ± 0.0131	1	0.5000	0	0.0000	0.0328	72.0091
23	0.78	79.79	98	72	0.7347 ± 0.0446	54	0.7500	59	0.8194	0.0306	82.6812
24	0.10	1542.58	91	83	0.9121 ± 0.0297	67	0.8072	29	0.3494	0.0205	705.8394

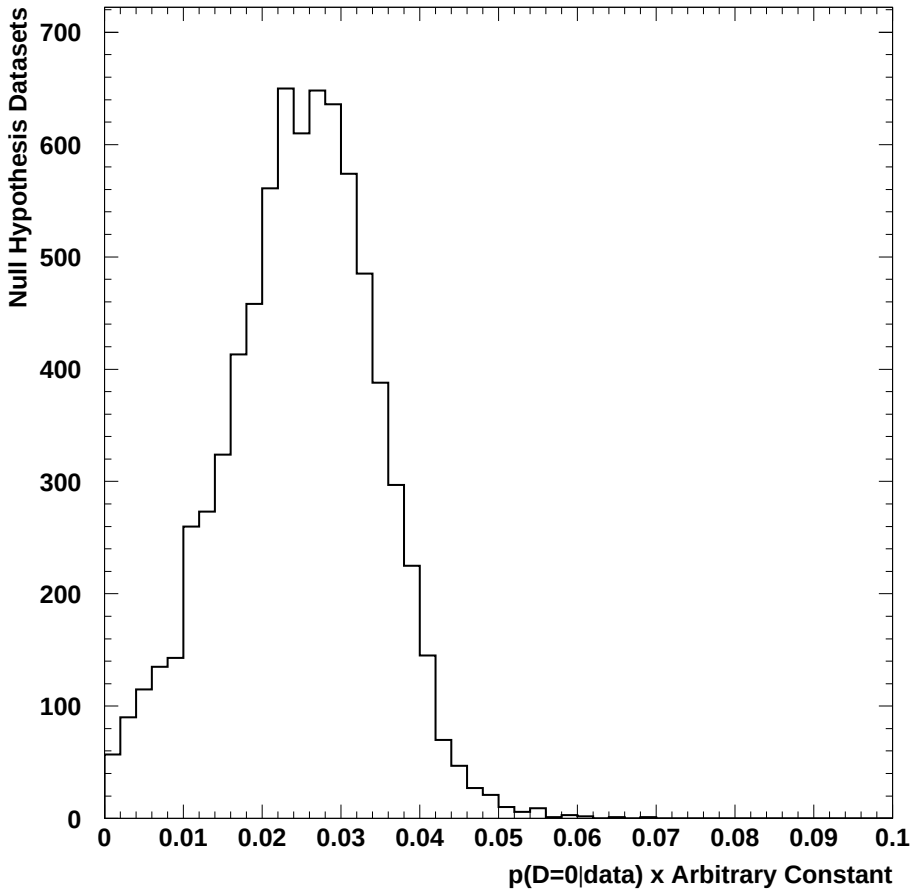

Figure 42: Distribution of Stefano's $p(D = 0|\text{data})$ value, up to a fixed multiplicative factor, in null hypothesis Problem 1 challenge datasets. Stefano places cuts of 3×10^{-3} and 4×10^{-3} for his two solutions to Problem 1.

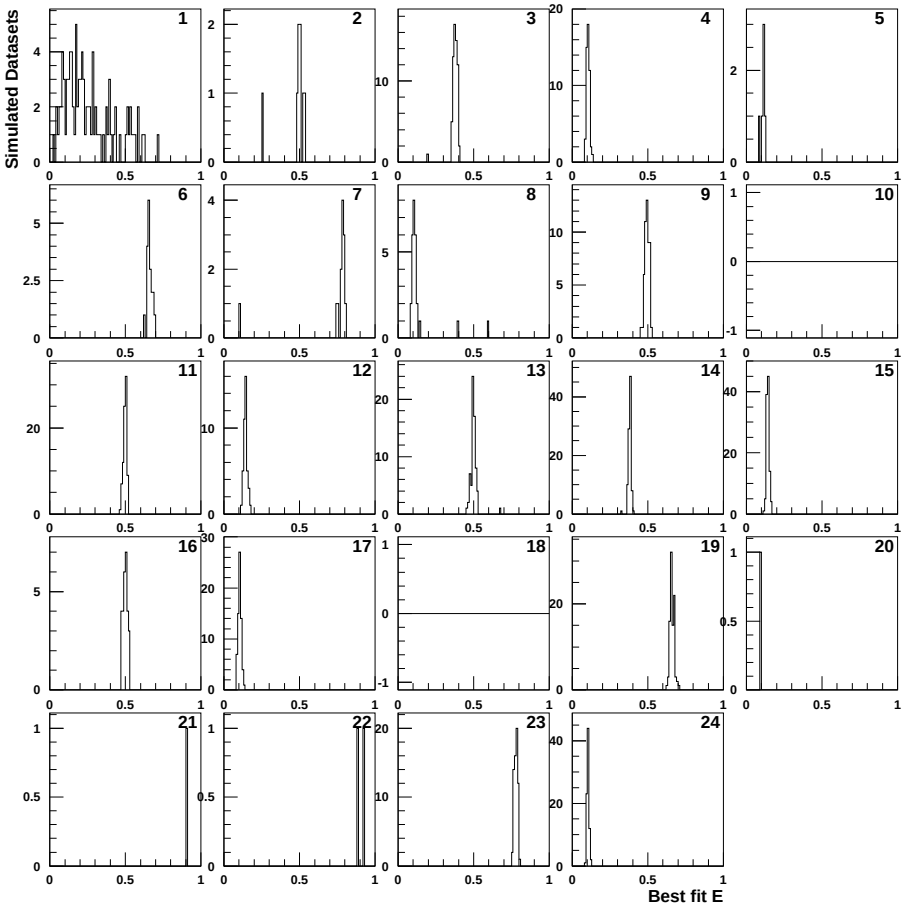

Figure 43: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Stefano claims evidence, split up by signal test category, for his solution with $p(D = 0|\text{data}) < 3 \times 10^{-3}$. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

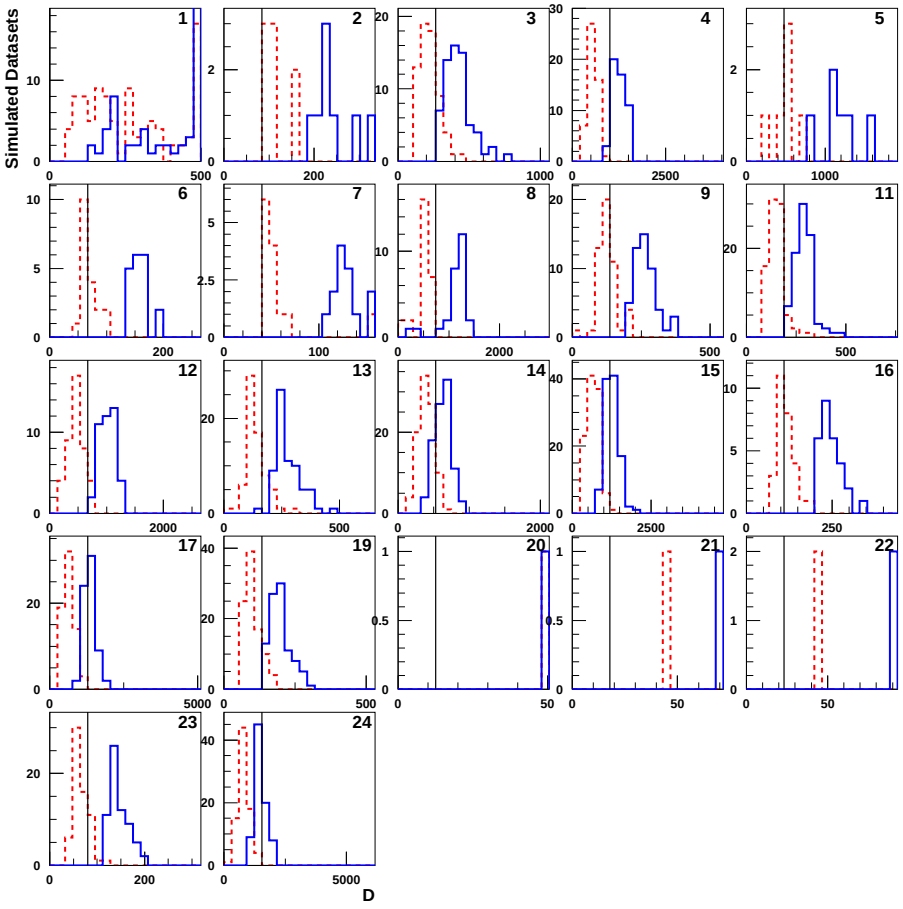

Figure 44: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Stefano claims evidence, split up by signal test category, for his solution with $p(D = 0|\text{data}) < 3 \times 10^{-3}$. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

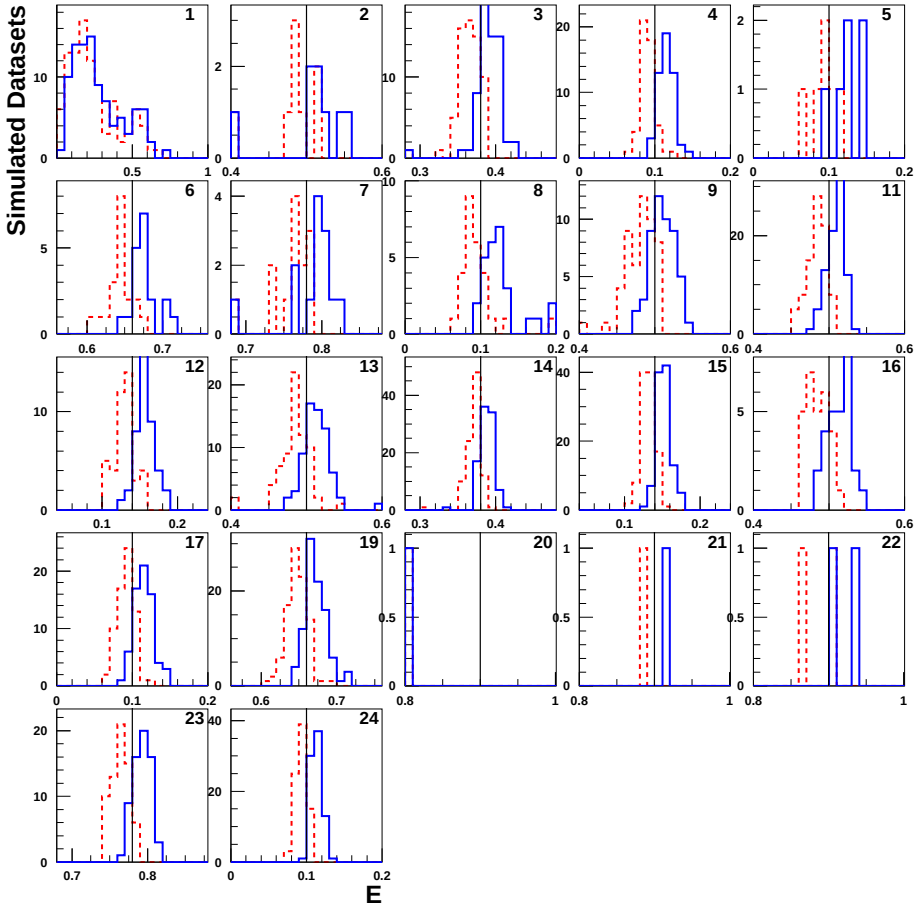

Figure 45: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Stefano claims evidence, split up by signal test category, for his solution with $p(D = 0|\text{data}) < 3 \times 10^{-3}$. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

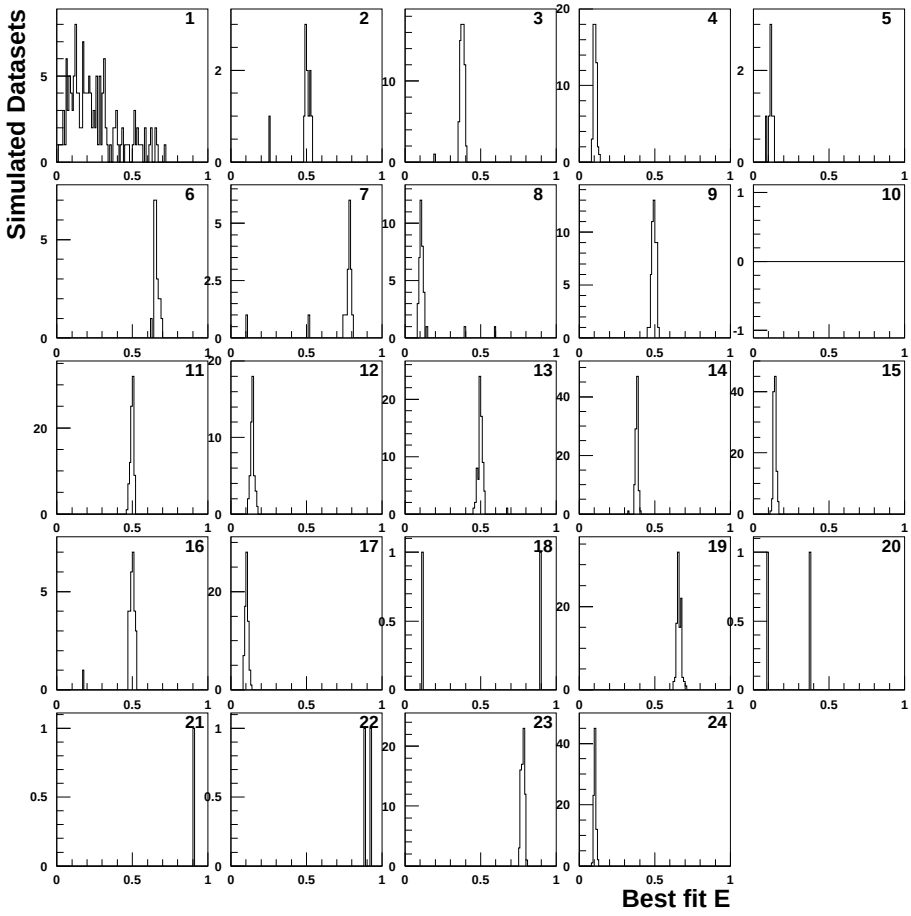

Figure 46: Distributions of the best fit values of E, the signal peak position in Problem 1, for challenge datasets for which Stefano claims evidence, split up by signal test category, for his solution with $p(D = 0|\text{data}) < 4 \times 10^{-3}$. The categories are listed in Table 1; the first category corresponds to the null hypothesis.

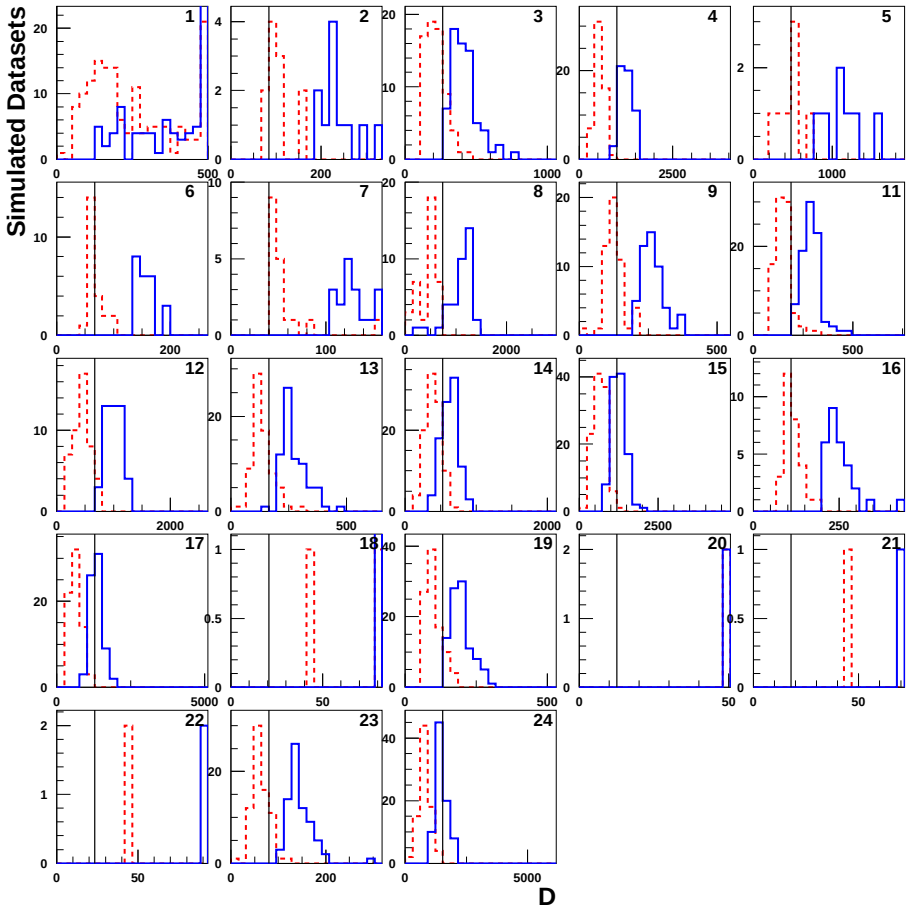

Figure 47: Distributions of the upper and lower interval edges for D, the signal rate parameter for Problem 1, for challenge datasets for which Stefano claims evidence, split up by signal test category, for his solution with $p(D = 0|\text{data}) < 4 \times 10^{-3}$. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

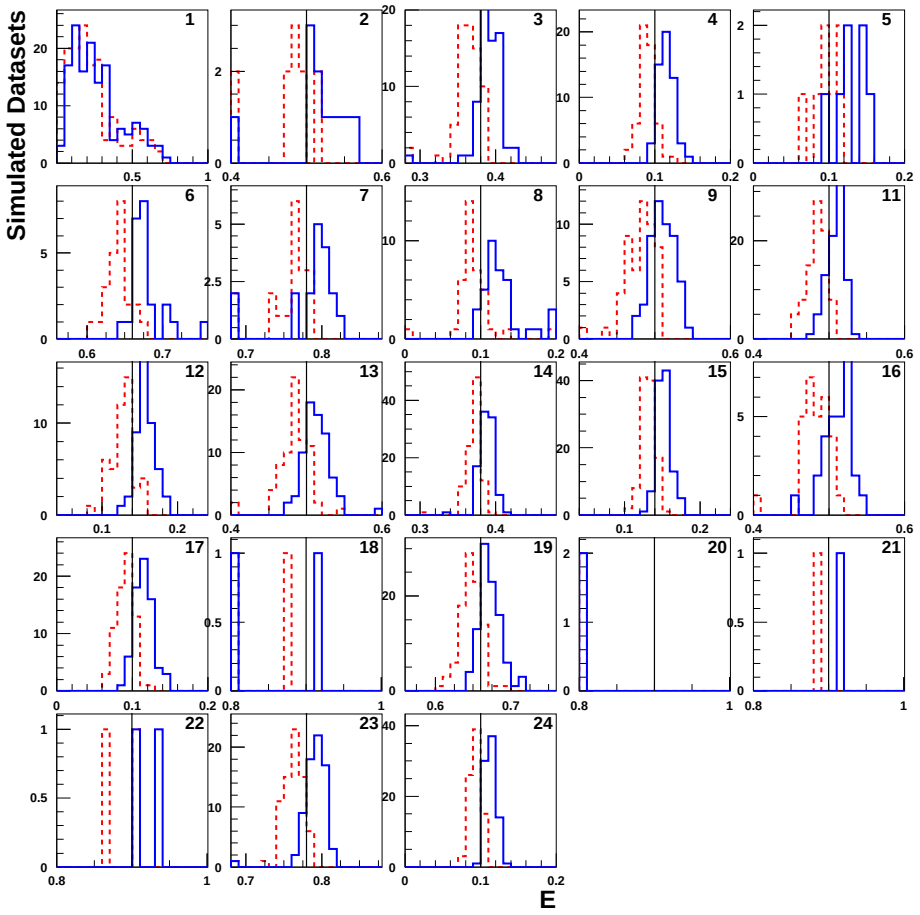

Figure 48: Distributions of the upper and lower interval edges for E, the signal position parameter for Problem 1, for challenge datasets for which Stefano claims evidence, split up by signal test category, for his solution with $p(D = 0|\text{data}) < 4 \times 10^{-3}$. The categories are listed in Table 1; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position.

2.3 Performance Summary Plots

The authors would like to thank Ofer Vitells, who collected the discovery fractions and coverages for the D and E parameters into a series of plots, labeled by the participant. These are shown in Figure 49, and are the discovery probabilities for each of the signal hypotheses, including the null hypothesis (D=0). The first graph shows the fraction of simulated datasets a discovery is claimed, and the datasets are ordered by the average fraction, averaged over the participants. The second graph shows the fraction of simulated datasets in which the injected signal rate parameter D lies within the 68% intervals quoted by the participants, separately for each of the signal models. The third graph shows the fraction of simulated datasets in which the peak position parameter E lies within the quoted intervals. The signal model number on the horizontal axes of these graphs is the same for each of the three graphs, but differs from that in Table 1.

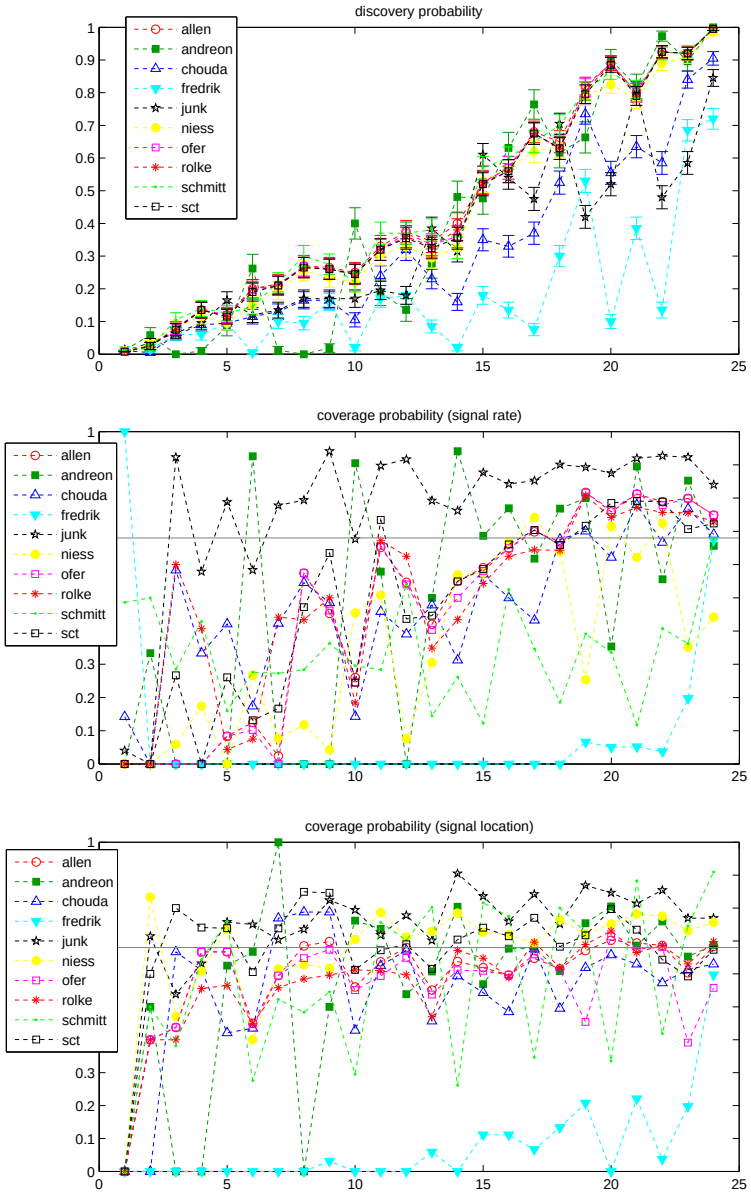

Figure 49: Top graph: The fraction of simulated datasets for which a discovery is claimed, separately for the null and test hypotheses. The hypothesis index on the abscissa is sorted by the average discovery probability. Middle and lower graphs: coverage fractions for the D and E parameters, respectively, with the same abscissa as the top graph.

Table 15: Problem 2 dataset categories – signal rates and how many repetitions of each were represented in the challenge datasets.

Category #	Input Signal	n_{rep}
1	0.00	17600
2	75.00	400
3	50.00	400
4	25.00	400
5	100.00	400
6	150.00	400
7	125.00	400

3 Challenge Problem #2

3.1 Simulated Datasets

Unlike Problem 1, Problem 2 parameterizes the predictions of the signal and background yields using finite samples of Monte Carlo. In a real high-energy physics experiment, sometimes samples of collider data are used instead. From a statistical standpoint, these are very similar and are treated identically. Often there is an extrapolation uncertainty associated with using a different sample of data which pass different selection requirements, and which are used to predict the background in the sample passing the signal requirements, and Monte Carlos are similarly fraught with uncertainty in their predictions.

Nonetheless, the simulated datasets and the Monte Carlo samples were in fact generated from smooth distributions for the marks. The distribution of the marks for Background 1 are given by

$$x = \max(1.0, 1.4y^{2.74}e^{-y/3}), \tag{2}$$

where y is uniformly distributed on the interval $(0, 1]$. Background 2 was generated with a uniform distribution. The signal distribution was generated using

$$x = z^{0.21}, \tag{3}$$

where z is uniformly distributed on the interval $(0, 1]$.

In each of the challenge datasets, a rate was chosen for Background 1, Background 2, and the signal, based on the hypothesis under test. The seven hypothesis categories are listed in Table 15. A Poisson random number was chosen using the randomly chosen rates, and then marks were generated using the prescriptions described above. The resulting lists of marks were then shuffled. The list of which simulated dataset was drawn from which signal test case was also shuffled.

Table 16: Listing of the Type-I error rates, and the estimated and measured correct-discovery rates for the three scenarios of Problem 2. Stefan Schmitt states that the power of his 50-bin test is similar to that of his 25-bin test.

Contributor	Type-I Error Rate	Signal = 75 Events	
	Measured	Claimed	Measured
Tom Junk	0.0068 ± 0.0006	0.865	0.870 ± 0.017
Wolfgang Rolke	0.0256 ± 0.0012	0.88	0.8500 ± 0.018
Stanford Challenge Team	0.0389 ± 0.0015	0.84	0.9100 ± 0.0143
Eilam Gross & Ofer Vitells	0.0107 ± 0.0008	0.815	0.7725 ± 0.0210
Valentin Niess	0.0085 ± 0.0007	0.761 ± 0.001	0.7125 ± 0.0226
Stefan Schmitt			
25 Bins	0.0047 ± 0.0005	0.85	0.8200 ± 0.0192
50 Bins	0.0047 ± 0.0005		0.8250 ± 0.0190
Doug Applegate & Matt Bellis	0.0168 ± 0.0010	0.95	0.8950 ± 0.0153

3.2 Solutions Received

Table 16 lists the contributors who provides solutions to Problem 2, the fractions of null-hypothesis simulated datasets that resulted in a discovery claim, and the fractions of simulated datasets that were in the power test samples that resulted in discovery claims, compared with the estimations provided by the participants.

3.2.1 From Tom Junk

Tom provided a solution to Problem 2 using a binned likelihood technique. Aside from the binning, and the lack of a peak position parameter, the method used is very similar to the solution used for Problem 1. An additional feature is the limited sample size of the Monte Carlo used to predict backgrounds. This adds an extra nuisance parameter for each bin for each sample – signal, background 1, and background 2. Tom fluctuates all of the nuisance parameters in each of his simulated data samples used to characterize the test statistic. This differs from the prior used to generate the datasets in that the characterizing datasets are binned, and the priors in each bin are taken as Gaussian approximations to the distributions of the bin-by-bin parameters. A possibly better choice is to use a Gamma prior in each bin for the bin-by-bin uncertainties, which is the Bayesian result using the finite Monte Carlo and a uniform prior in the unknown true background and signal rates. This however biases up the prediction in each bin. Tom fit the two background rates, but did not fit the separate bin-by-bin uncertainties, to get values of the $-2\ln Q$ test statistic for the simulated datasets and the challenge datasets.

For the signal rate intervals, Tom performed a Bayesian calculation, integrating the likelihood function times a uniform prior in the signal rate over the uncertain parameters (this time, the two background rates and the bin-by-bin uncertainties). The 68% credibility interval is computed as the shortest interval containing 68% of the integral of the posterior.

Table 17 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data. Tom's Type-I error rate is $(0.68 \pm 0.06)\%$, computed on simulated datasets with no signal present. The fit signal rate measurements also all cover the true signal rate at more than the 68% level, except for the smallest signal case, in which all simulated datasets which had a sufficiently small p value gave too large a fitted cross section. The problem may have been better formulated if cross section measurements were requested even if evidence for a signal is not claimed, although this situation does not arise frequently in particle physics. Thus there is a natural bias towards larger signal rate fits for experiments reporting evidence of new signals, as those experiments failing to obtain evidence do not publish cross sections.

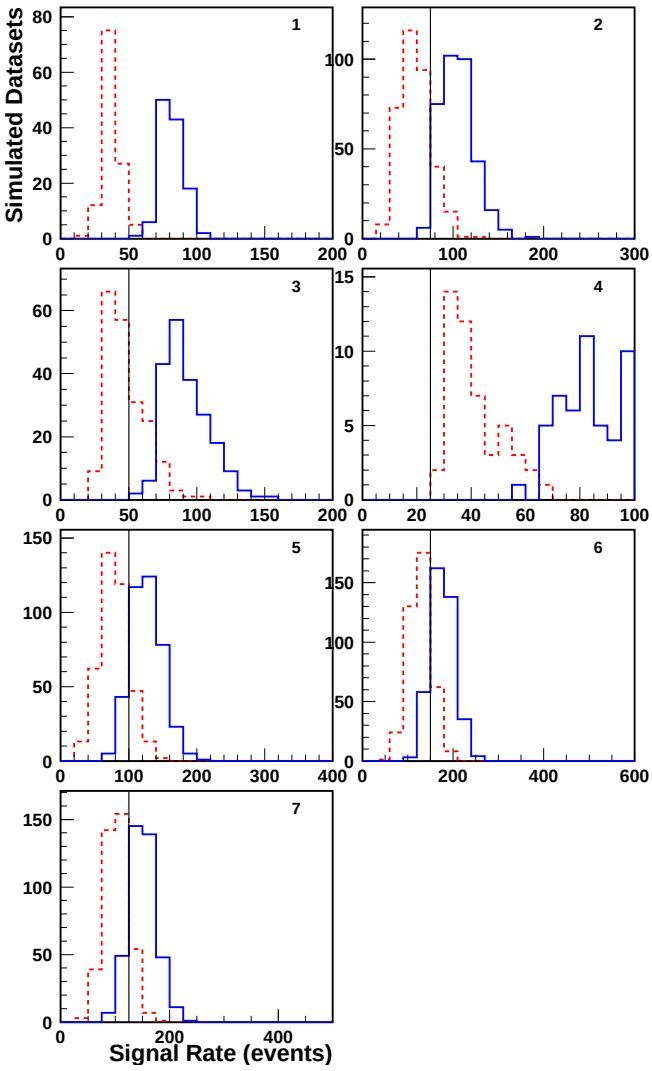

Figure 50: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Tom claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

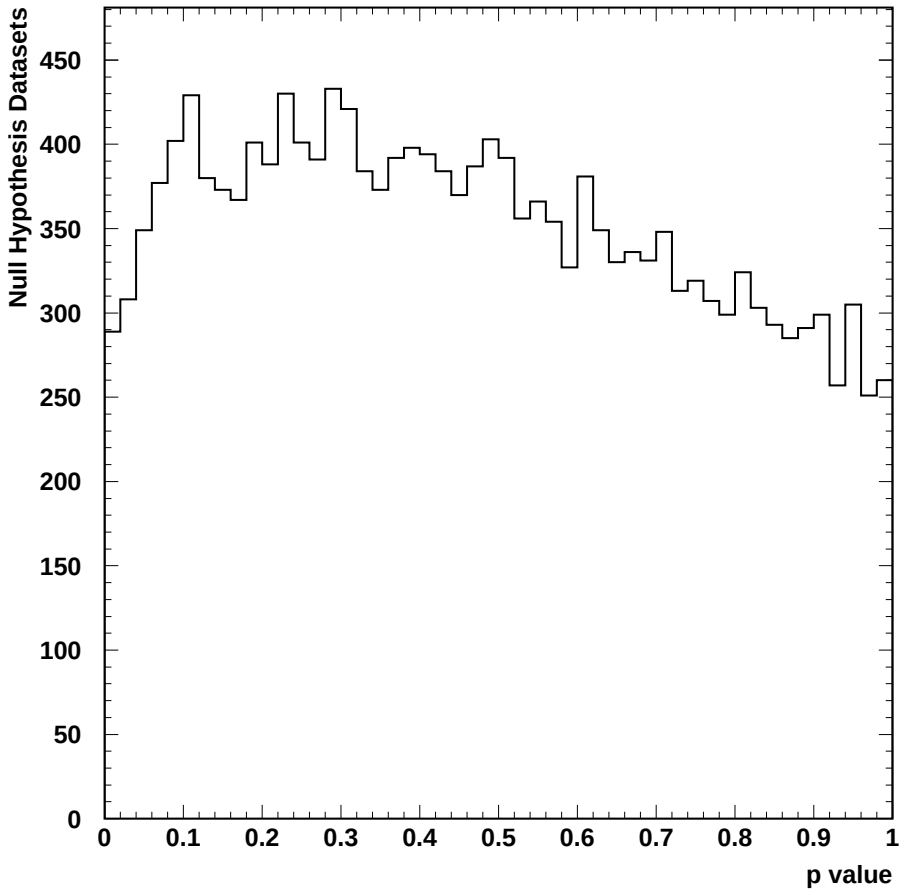

Figure 51: Distribution of the quoted p value in null hypothesis challenge datasets for Tom's solution to Problem 2.

Table 17: Problem 2 performance evaluation for Tom Junk's solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	$Sig_{\rm true}$	$n_{\rm rep}$	$n_{\rm disc}$	$f_{\rm disc}$	$n_{\rm Scorr}$	$f_{\rm Scorr}$	$\langle$Swid$\rangle$
1	0.00	17600	120	0.0068 ± 0.0006	0	0.0000	44.7819
2	75.00	400	348	0.8700 ± 0.0168	286	0.8218	45.9750
3	50.00	400	205	0.5125 ± 0.0250	132	0.6439	44.5449
4	25.00	400	49	0.1225 ± 0.0164	0	0.0000	42.9888
5	100.00	400	396	0.9900 ± 0.0050	286	0.7222	47.5670
6	150.00	400	400	1.0000 ± 0.0000	270	0.6750	50.4234
7	125.00	400	400	1.0000 ± 0.0000	282	0.7050	48.9148

Table 18: Problem 2 performance evaluation for Wolfgang Rolke's solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle$Swid$\rangle$
1	0.00	17600	451	0.0256 ± 0.0012	0	0.0000	44.8011
2	75.00	400	340	0.8500 ± 0.0179	216	0.6353	49.4226
3	50.00	400	235	0.5875 ± 0.0246	100	0.4255	47.4689
4	25.00	400	90	0.2250 ± 0.0209	0	0.0000	46.4322
5	100.00	400	395	0.9875 ± 0.0056	256	0.6481	50.9408
6	150.00	400	400	1.0000 ± 0.0000	225	0.5625	53.5235
7	125.00	400	399	0.9975 ± 0.0025	240	0.6015	52.4113

3.2.2 From Wolfgang Rolke

Wolfgang Rolke provided a solution to Problem 2 using a likelihood ratio test similar to that used in Problem 1. Since the background and signal predictions are Monte Carlo based, Wolfgang tried a parametric description, fitting Beta functions to the signal and background shapes, and a non-parametric description, which bins the results. Semi–parametric solutions are also possible, in which some components are parameterized and others are binned. Wolfgang found very similar performance for the the nonparametric and parametric approaches, and he chooses his parametric solution. Background 1 requires a little effort to get the low end and the high end to fit well, as there is so much of it populating the low end. The distribution of the log likelihood ratio λ is fit to a χ^2 distribution with one degree of freedom. We expect this since the test hypothesis has one extra free parameter, the signal rate, compared with the null hypothesis.

Table 18 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data. Wolfgang's Type-I error rate is $(2.56 \pm 0.12)\%$, computed on simulated datasets with no signal present. We can see that the p value distribution rises at low p values, consistent with the larger-than-desired Type-I error rate.

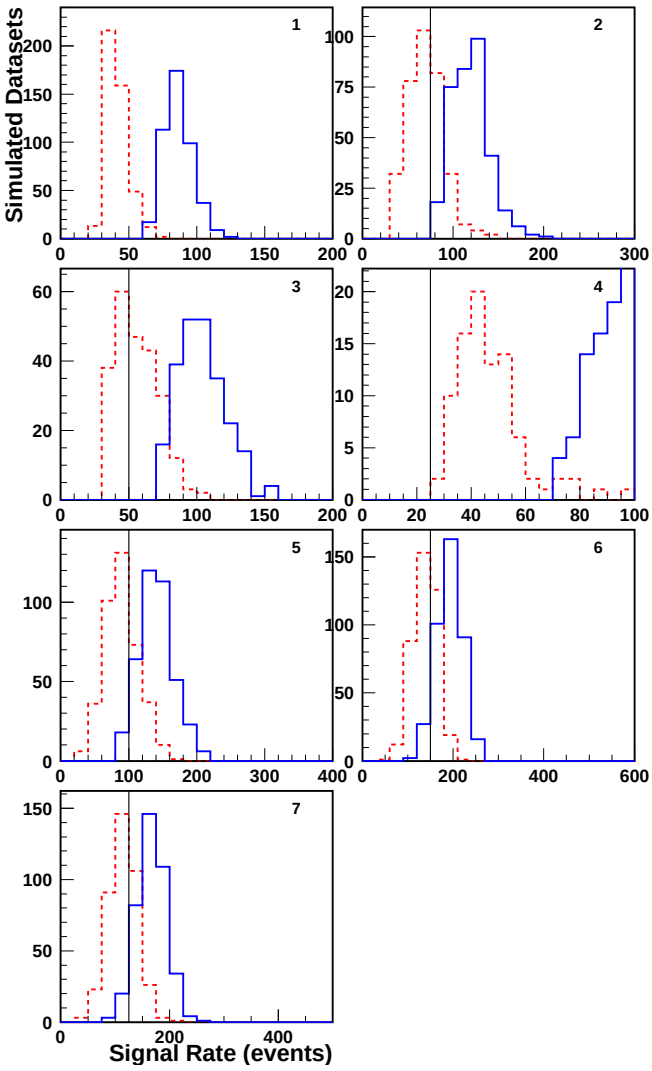

Figure 52: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Wolfgang claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

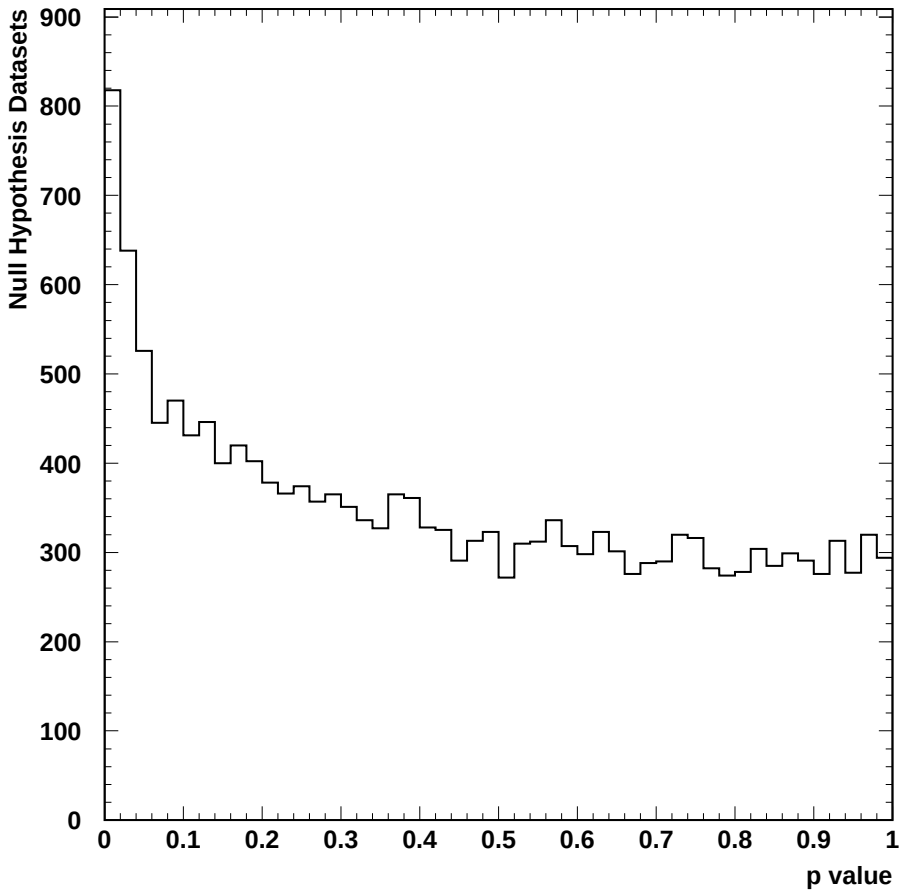

Figure 53: Distribution of the quoted p value in null hypothesis challenge datasets for Wolfgang's solution to Problem 2.

Table 19: Problem 2 performance evaluation for the SCT's solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle$Swid$\rangle$
1	0.00	17600	685	0.0389 ± 0.0015	0	0.0000	27.4298
2	75.00	400	364	0.9100 ± 0.0143	181	0.4973	32.4090
3	50.00	400	265	0.6625 ± 0.0236	102	0.3849	30.4449
4	25.00	400	116	0.2900 ± 0.0227	4	0.0345	28.6076
5	100.00	400	397	0.9925 ± 0.0043	204	0.5139	34.8550
6	150.00	400	400	1.0000 ± 0.0000	193	0.4825	38.8669
7	125.00	400	399	0.9975 ± 0.0025	213	0.5338	36.9672

3.2.3 From the Stanford Challenge Team

The SCT provided a solution to Problem 2 using a likelihood ratio test similar to that used in Problem 1, comparing a three-component fit to a two-component fit (three including the signal, and two backgrounds are fit in either hypothesis). The distributions of the marks for the two background components and the signal component are approximated with Beta distributions.

Table 19 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data. The SCT's Type-I error rate is 0.0389 ± 0.0015, computed on simulated datasets with no signal present, well in excess of the desired 1%. Two interesting features of the p value distribution, shown in Figure 55, are that the p value never exceeds 0.5 (the top half of the distribution appears to be concentrated at $p = 0.5$), and that the distribution of p rises at low p, thus causing concern for the lack of coverage.

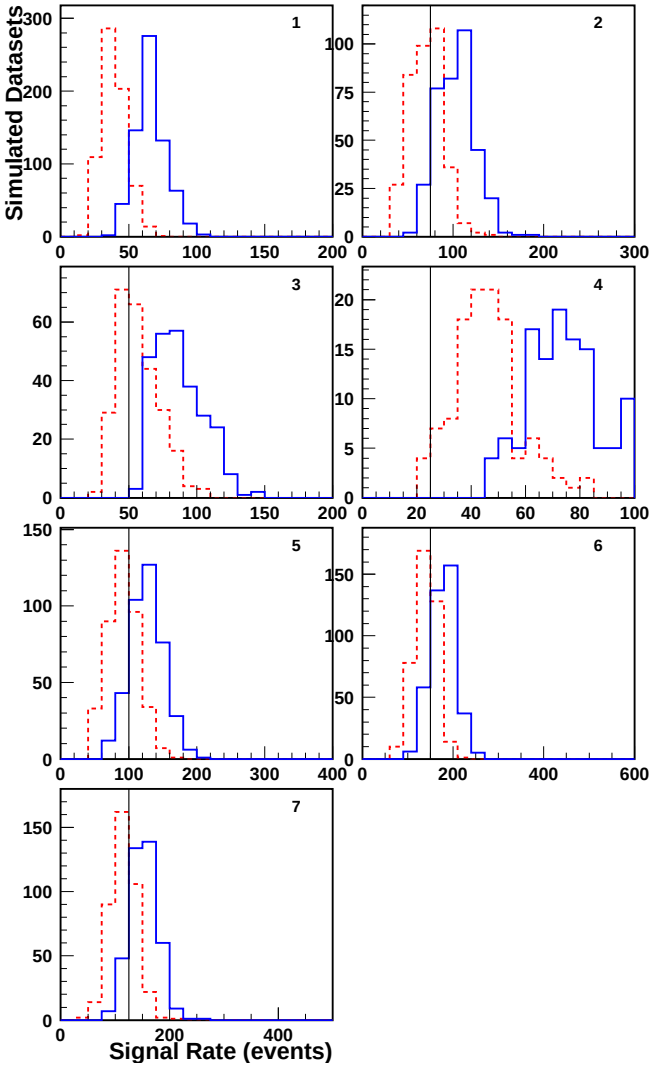

Figure 54: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which the SCT claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

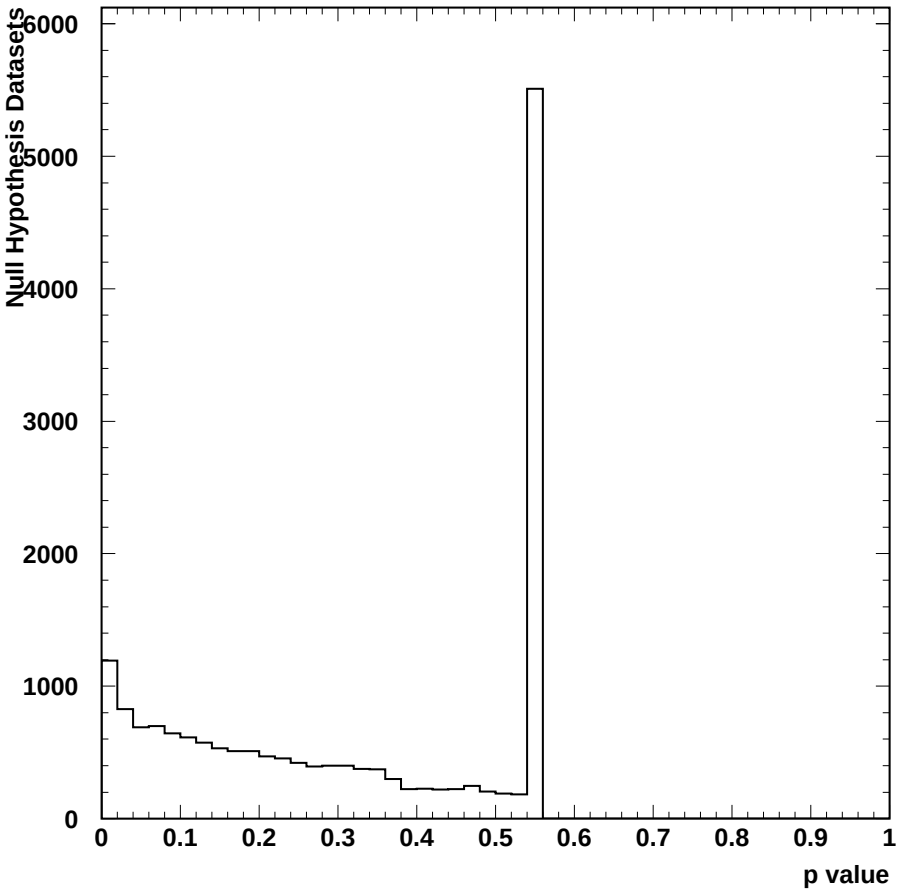

Figure 55: Distribution of the quoted p value in null hypothesis challenge datasets for the SCT's solution to Problem 2.

Table 20: Problem 2 performance evaluation for Eilam and Ofer's solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle \text{Swid} \rangle$
1	0.00	17600	188	0.0107 ± 0.0008	0	0.0000	44.5158
2	75.00	400	309	0.7725 ± 0.0210	262	0.8479	49.1210
3	50.00	400	182	0.4550 ± 0.0249	109	0.5989	47.2930
4	25.00	400	54	0.1350 ± 0.0171	0	0.0000	45.3404
5	100.00	400	382	0.9550 ± 0.0104	276	0.7225	50.8607
6	150.00	400	400	1.0000 ± 0.0000	279	0.6975	54.0924
7	125.00	400	397	0.9925 ± 0.0043	278	0.7003	52.4133

3.2.4 From Eilam Gross and Ofer Vitells

Eilam and Ofer provided a solution to Problem 2 using a likelihood ratio test statistic similar to that of Problem 1, except in this case the likelihood ratio is binned, and there is no Look Elsewhere Effect.

Table 20 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data. Eilam and Ofer's Type-I error rate is 0.0107 ± 0.0008 computed on simulated datasets with no signal present, which is not measurably different from the desired value of 1%.

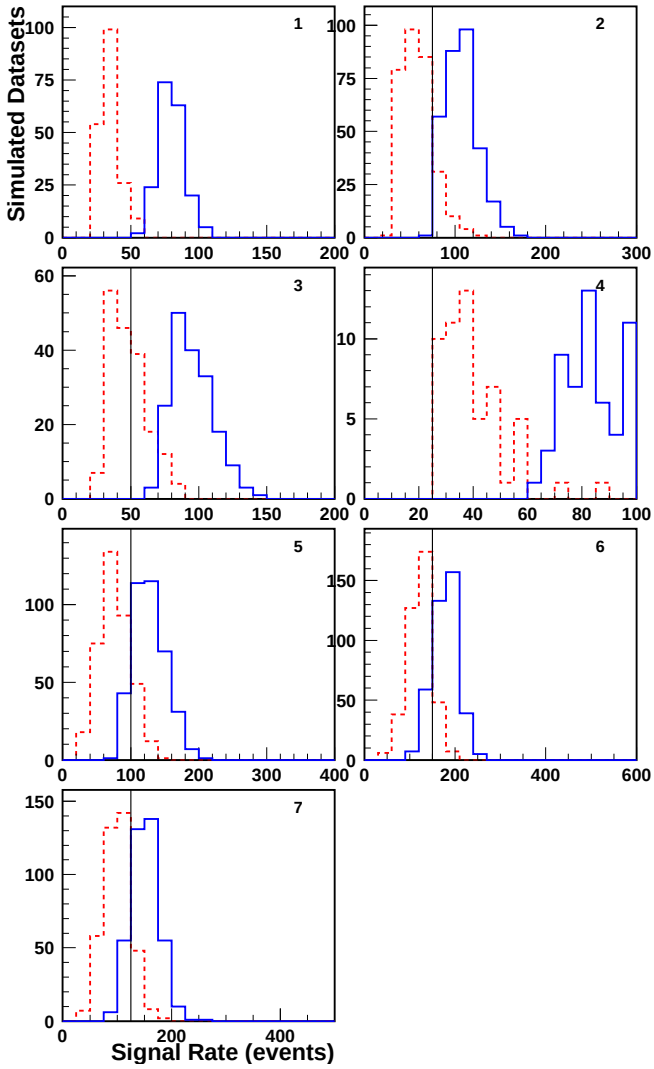

Figure 56: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Eilam and Ofer claim evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

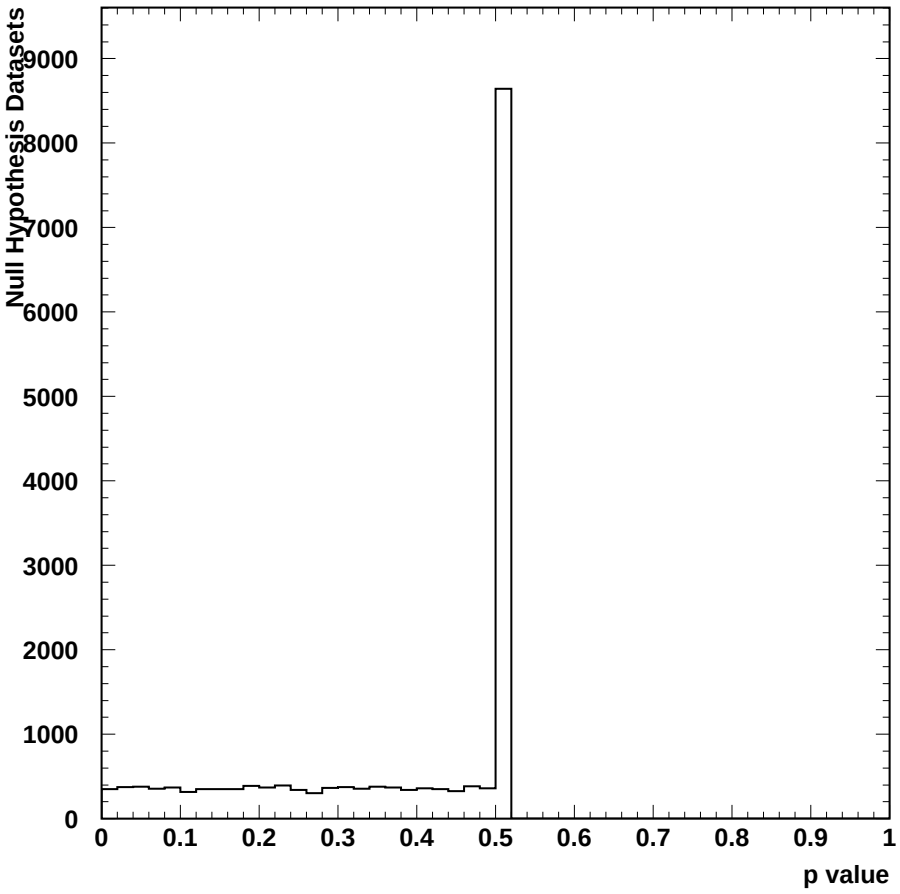

Figure 57: Distribution of the quoted p value in null hypothesis challenge datasets for the Eilam and Ofer's solution to Problem 2.

Table 21: Problem 2 performance evaluation for Valentin Niess's solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle$Swid$\rangle$
1	0.00	17600	150	0.0085 ± 0.0007	0	0.0000	42.4200
2	75.00	400	285	0.7125 ± 0.0226	193	0.6772	52.3930
3	50.00	400	156	0.3900 ± 0.0244	55	0.3526	47.5449
4	25.00	400	47	0.1175 ± 0.0161	0	0.0000	41.5106
5	100.00	400	366	0.9150 ± 0.0139	261	0.7131	54.7268
6	150.00	400	400	1.0000 ± 0.0000	270	0.6750	54.9250
7	125.00	400	391	0.9775 ± 0.0074	269	0.6880	54.7442

3.2.5 From Valentin Niess

Valentin Niess provided a solution to Problem 2 using a Kolmogorov-Smirnov test, parameterizing the signal and background cumulative distributions with power-law functions of the marks. The KS test statistic is minimized over the uncertain values of the signal and background rates numerically using the PORT library.

Table 21 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data. Valentin's Type-I error rate is 0.0085 ± 0.0007 computed on simulated datasets with no signal present, which is comfortably less than 1%.

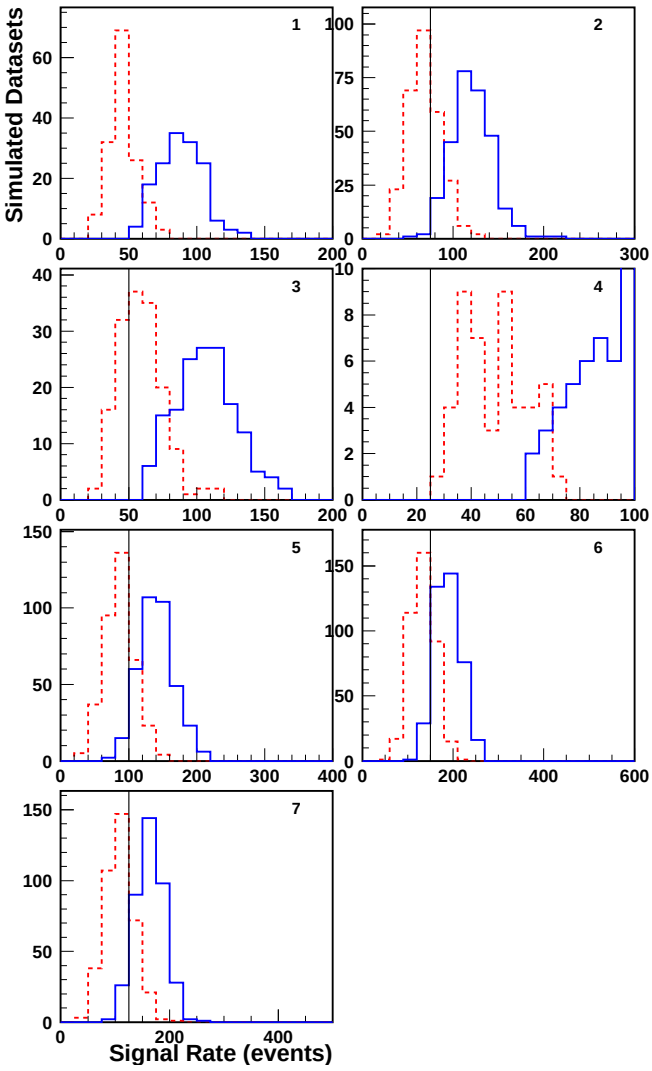

Figure 58: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Valentin claims evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

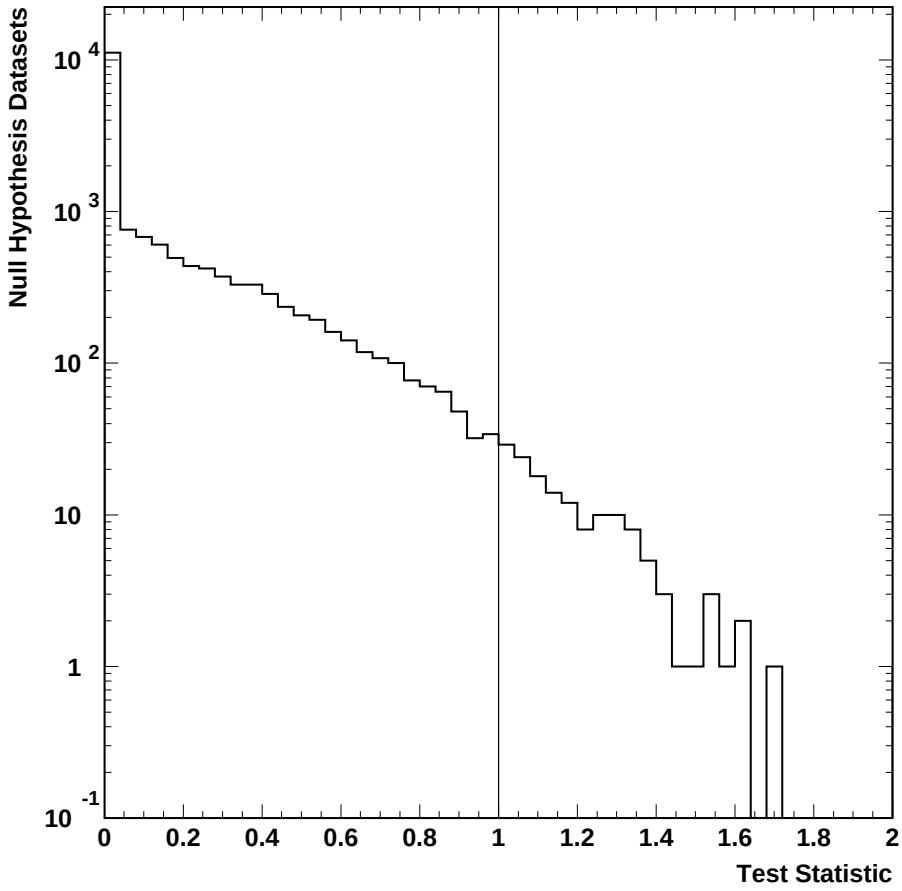

Figure 59: Distribution of the quoted test statistic value divided by the critical value in null hypothesis challenge datasets for Valentin's solution to Problem 2.

Table 22: Problem 2 performance evaluation for Stefan Schmitt's 25-bin solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle$Swid$\rangle$
1	0.00	17600	82	0.0047 ± 0.0005	2	0.0244	56.3457
2	75.00	400	328	0.8200 ± 0.0192	248	0.7561	51.3742
3	50.00	400	183	0.4575 ± 0.0249	140	0.7650	53.1921
4	25.00	400	35	0.0875 ± 0.0141	22	0.6286	54.5959
5	100.00	400	394	0.9850 ± 0.0061	284	0.7208	50.8824
6	150.00	400	400	1.0000 ± 0.0000	280	0.7000	53.1749
7	125.00	400	400	1.0000 ± 0.0000	288	0.7200	51.5327

Table 23: Problem 2 performance evaluation for Stefan Schmitt's 50-bin solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle$Swid$\rangle$
1	0.00	17600	82	0.0047 ± 0.0005	2	0.0244	55.9555
2	75.00	400	330	0.8250 ± 0.0190	245	0.7424	51.2024
3	50.00	400	175	0.4375 ± 0.0248	133	0.7600	52.7358
4	25.00	400	34	0.0850 ± 0.0139	22	0.6471	54.5620
5	100.00	400	395	0.9875 ± 0.0056	278	0.7038	50.5337
6	150.00	400	400	1.0000 ± 0.0000	277	0.6925	52.7789
7	125.00	400	400	1.0000 ± 0.0000	287	0.7175	51.1493

3.2.6 From Stefan Schmitt

Stefan Schmitt provided a solution to Problem 2 using a weighted event-counting technique. Two solutions were provided, one choosing 25 bins for the marks and the other choosing 50 bins.

Table 22 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data, for the 25-bin solution, and Table 23 lists the same information for Stefan's 50-bin solution.

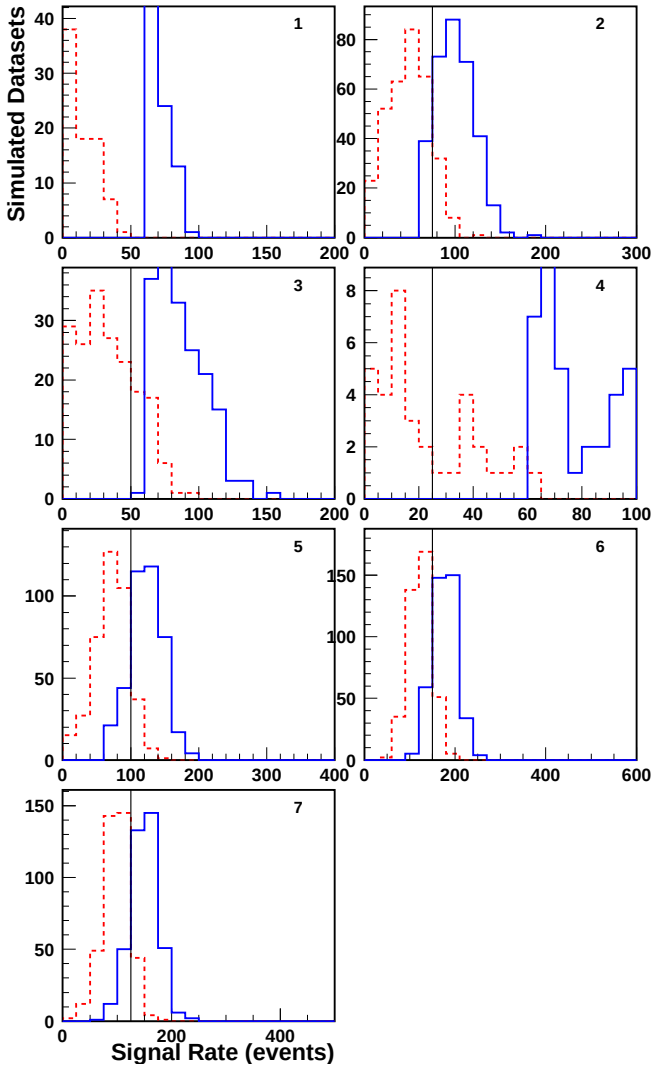

Figure 60: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his 25-bin solution. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

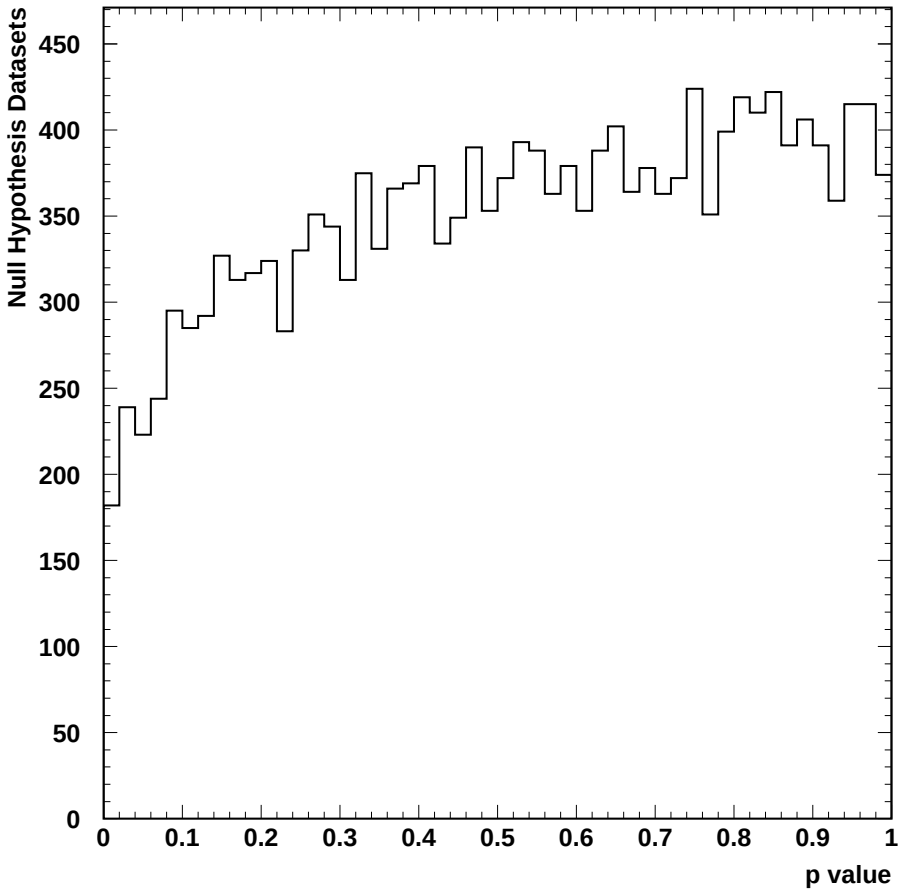

Figure 61: Distribution of the quoted p value in null hypothesis challenge datasets for Stefan's 25-bin solution to Problem 2.

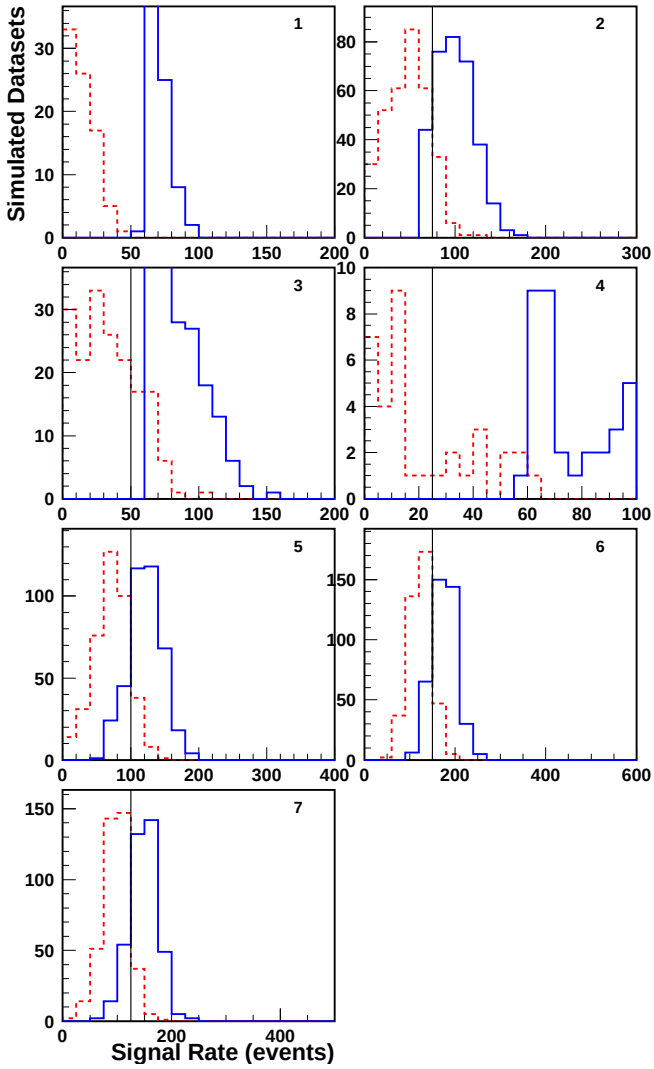

Figure 62: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Stefan claims evidence with a claimed Type-I error rate of 1%, split up by signal test category, for his 50-bin solution. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

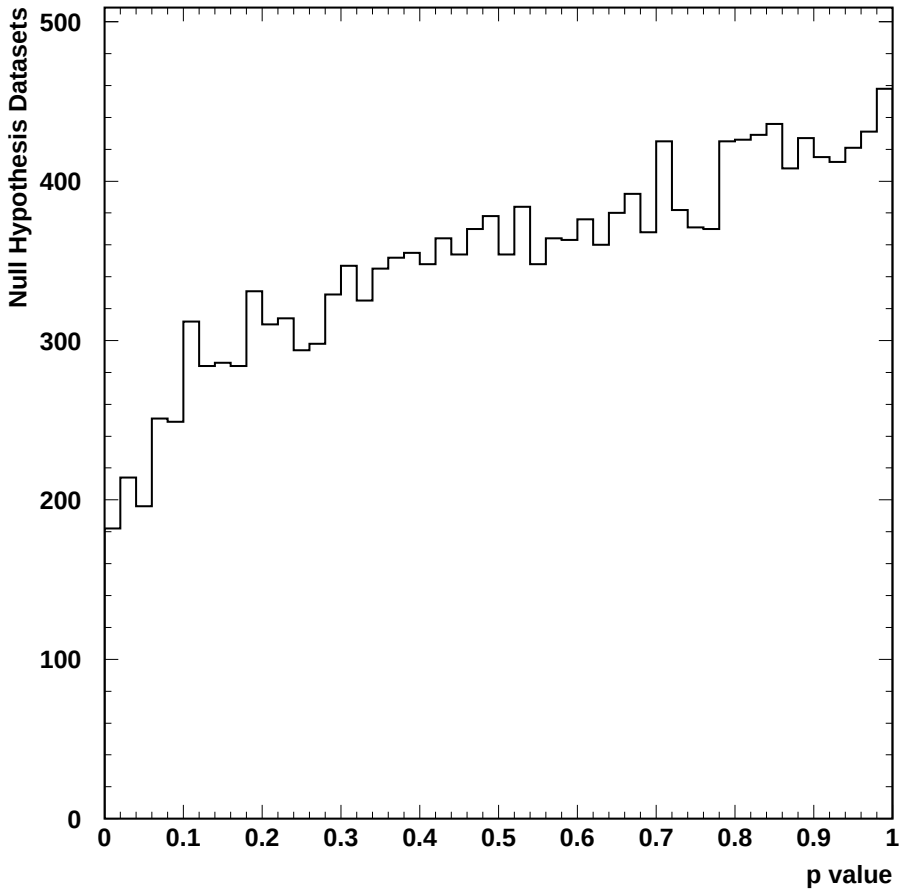

Figure 63: Distribution of the quoted p value in null hypothesis challenge datasets for Stefan's 50-bin solution to Problem 2.

Table 24: Problem 2 performance evaluation for Doug and Matt's solution, showing the Type-I and 1-Type-II error rates. The categories are listed in Table 15.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle$Swid$\rangle$
1	0.00	17600	295	0.0168 ± 0.0010	22	0.0746	87.6877
2	75.00	400	358	0.8950 ± 0.0153	272	0.7598	74.3316
3	50.00	400	244	0.6100 ± 0.0244	154	0.6311	76.0376
4	25.00	400	81	0.2025 ± 0.0201	7	0.0864	71.6980
5	100.00	400	398	0.9950 ± 0.0035	291	0.7312	68.6577
6	150.00	400	400	1.0000 ± 0.0000	259	0.6475	77.7485
7	125.00	400	400	1.0000 ± 0.0000	286	0.7150	79.0725

3.2.7 From Matt Bellis

Doug Applegate and Matt Bellis provided a solution to Problem 2 using a nearest neighbor approach to classify events, giving them probability weights to have come from the three processes. The Monte Carlo samples are finite in size, and thus a bootstrap technique is used.

Table 24 lists the discovery rates for each of the seven scenarios embedded in Problem 2's test data. Doug and Matt's Type-I error rate is 0.0168 ± 0.0010, computed on simulated datasets with no signal present.

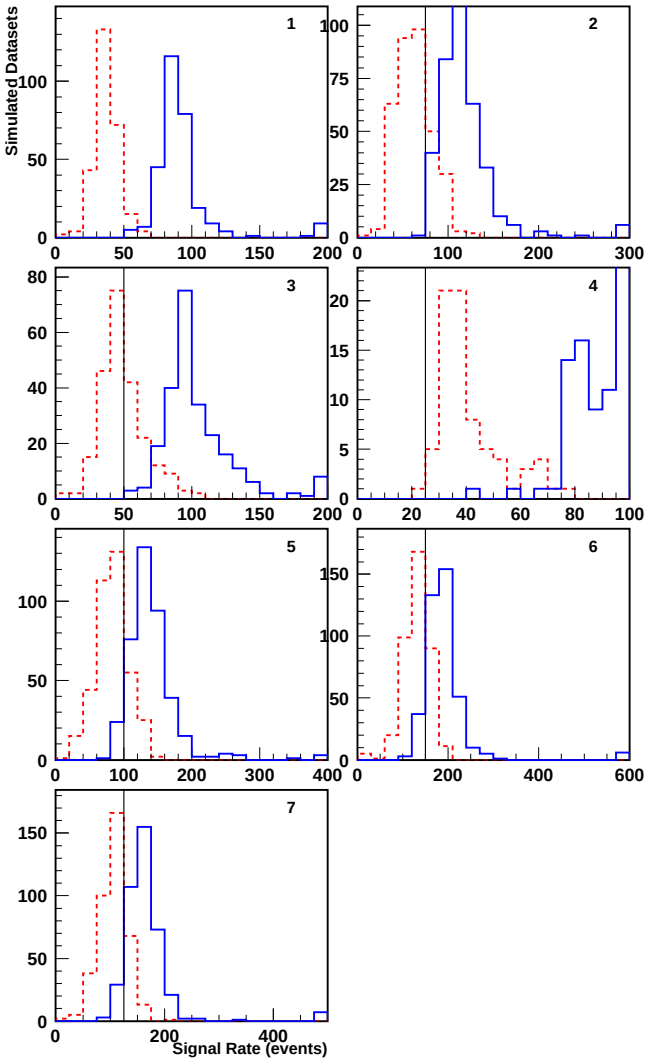

Figure 64: Distributions of the upper and lower interval edges for the signal rate for Problem 2, for challenge datasets for which Doug and Matt claim evidence with a claimed Type-I error rate of 1%, split up by signal test category. The categories are listed in Table 15; the first category corresponds to the null hypothesis. The red dashed histograms show the distributions of the lower edge of the reported intervals, and the blue solid histograms show the distribution of the upper edges. The black lines show the true signal position. Overflows are collected in the last bin. The vertical scales on this plot are set by the red dashed histograms – the blue histograms may extend above the tops of the panels.

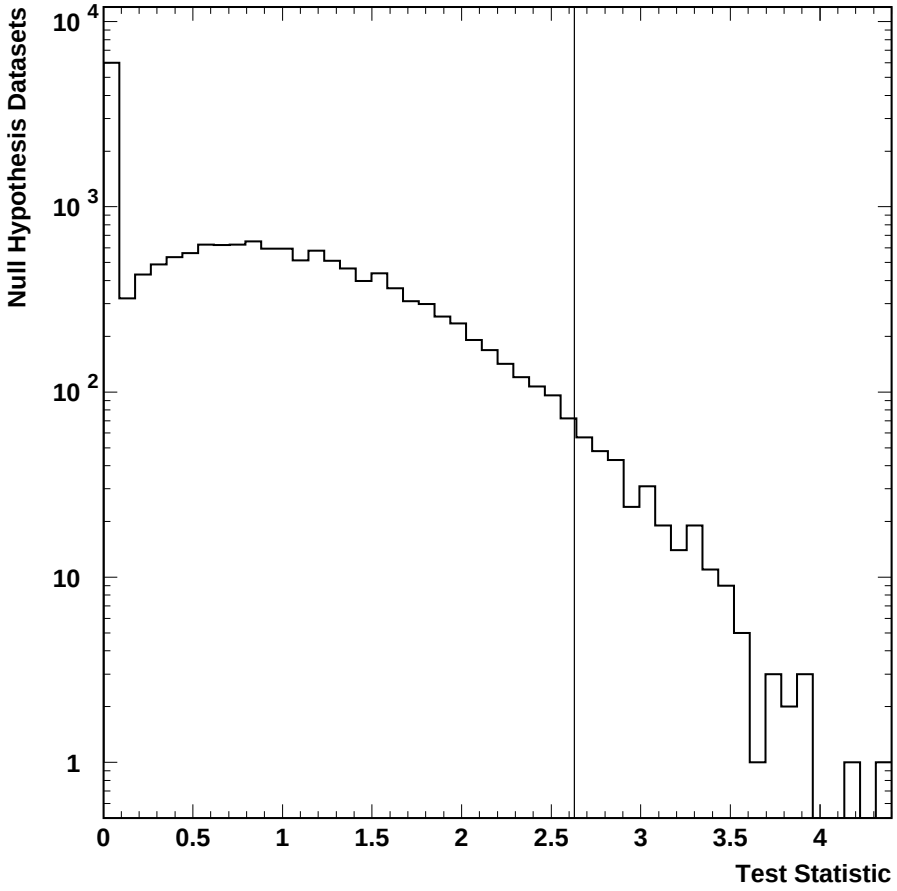

Figure 65: Distribution of the quoted test statistic in null hypothesis challenge datasets for Doug and Matt's solution to Problem 2.

4 Summary

The solutions to the Banff Challenge 2a problems provided by the participants spanned a range of different approaches. Most of the hypothesis tests were based on a ratio of profile likelihoods, with Monte Carlo simulation of the the distribution of the test statistic. Minor variations between submissions arise from the choice of binning or unbinned fits, and the strategy used to find a global minimum among many local minima in the first problem, and in the parameterization and handling of the distributions of the marks in the second problem. Alternate approaches involved counting events inside signal windows while fitting backgrounds in the sidebands, counting fractional events, and using the Kolmogorov-Smirnov test statistic.

The Look-Elsewhere Effect is an issue in Problem 1 but not in Problem 2, since the presence of a signal introduces an additional parameter – the location of the peak E in the test hypothesis which is not present in the null hypothesis. All participants handled this effect rather well – there are no signs of noticeable undercoverage in the Type-I error rate measurements. One of the methods of accounting for the Look-Elsewhere Effect had the effect of producing p values in excess of unity however.

A typical particle physics experiment has a flip-flopping approach of when to quote a two-sided interval and when to quote a one-sided upper limit. The part of the challenge specification asking for two-sided intervals when evidence was claimed and otherwise not did not allow a unified approach, and also biased the intervals on the rate parameter upwards, most noticeably in the simulated datasets drawn from the null hypothesis. Quoting a two-sided interval for the production rate of a new particle for which evidence is not claimed can be misconstrued by the broader community, even though doing so would help the coverage properties of the methods.

It is in Problem 2 that significant undercoverage, that is, a higher-than-expected Type-I error rate, was seen in several submissions. Participants both underestimated their Type-I error rates and overestimated their discovery power. Because the distributions of the marks were not given to the participants, instead relying on simulated Monte Carlo samples of them, participants either binned the data or calculated unbinned likelihoods using parameterizations that appear to fit the distributions of the marks in the simulated Monte Carlo samples. It could also be that the *a priori* uncertainty of 100% on the rate of Background 2 causes ambiguities to arise in the approach to follow that is reflected measurably in the results, particularly since Background 2 looks more like the signal than Background 1 looks like the signal.

Acknowledgments

We would like to thank Ofer Vitells who made the performance summary plots for Problem 1.

References

[1] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics", Eur. Phys. J. C **70**, 1-2 (2010).

[2] G. Choudalakis, [arXiv:1101.0390 [physics.data-an]] (2011).